FUNDAMENTAL FREQUENCY ALIGNMENT VS. NOTE-BASED MELODIC SIMILARITY FOR SINGING VOICE ASSESSMENT

Emilio Molina[1], *Isabel Barbancho*[1], *Emilia Gómez*[2], *Ana María Barbancho*[1], *Lorenzo J. Tardón*[1]

[1]Dept. Ingeniería de Comunicaciones, ETSI Telecomunicación, Universidad de Málaga, Spain
[2]Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
emm@ic.uma.es, ibp@ic.uma.es, emilia.gomez@upf.edu, abp@ic.uma.es, lorenzo@ic.uma.es

ABSTRACT

This paper presents a generic approach for automatic singing assessment for basic singing levels. The system provides the user with a set of intonation, rhythm and overall ratings obtained by measuring the similarity of the sung melody and a target performance. Two different similarity approaches are discussed: f_0 curve alignment through Dynamic Time Warping (DTW), and singing transcription plus note-level similarity. From these two approaches, we extract different intonation and rhythm similarity measures which are combined through quadratic polynomial regression analysis in order to fit the judgement of 4 trained musicians on 27 performances. The results show that the proposed system is suitable for automatic singing voice rating and that DTW based measures are specially simple and effective for intonation and rhythm assessment.

Index Terms— singing assessment, automatic transcription, score alignment, melodic similarity, singing voice

1. INTRODUCTION

The assessment of a given musical performance is commonly affected by many subjective factors, even in the case of expert musicians [1]. Therefore, the development of an automatic performance evaluation system is a challenging problem. Under controlled conditions, some objective aspects can be considered and computationally modelled. Some studies have analyzed the reliability of judgements in music performance evaluation [1, 2, 3]. In such studies, different musicians were asked to rate a certain number of performers according to different aspects, with the aim of studying how objective the different judgements were. Some aspects such as intonation accuracy, vibrato or rhythm seem to be quite reliably judged by musicians, unlike more subjective aspects such as diction.

This work has been funded by the Ministerio de Economia y Competitividad of the Spanish Government under Project No. TIN2010-21089-C03-02 and Project No. IPT-2011-0885-430000 and by the Ministerio de Industria, Turismo y Comercio under Project No. TSI-090100-2011-25. This work has been partially done under Campus de Excelencia Internacional CEI Andalucía TECH in the context of Program Campus de Excelencia Internacional of the Spanish Ministerio de Educación.

Prior work has lead to various solutions for automatic singing rating [4, 5, 6, 7, 8, 9, 10]. In general, all these systems focus on intonation assessment with visually attractive real-time feedback. Songs2See [10] is a recent and representative example of the state of the art. Nevertheless, current approaches do not generally handle rhythmic misalignments, and the feedback provided is not directly based on trained musicians' judgements. This study deals with automatic intonation and rhythm assessment of singing performances, being our main goal to provide the user with meaningful feedback based on modeling teachers' criteria. We focus on basic singing levels, i.e. children and beginners. Two different approaches for singing assessment are evaluated: dynamic time warping (DTW) and note-level similarity with respect to a target melody.

This paper is organized as follows: Section 2 provides an overall description of the selected approach. The evaluation methodology is presented in Section 3, including ground truth gathering (Section 3.1) and evaluation measures (Section 3.2). Section 4 presents our main results and Section 5 draws some conclusions about this study.

2. SELECTED APPROACH

We propose a generic schema for singing assessment based on melodic similarity with respect to a target melody. The overall block diagram is illustrated in Figure 1. The audio input is first analyzed to extract a set of low-level descriptors (Section 2.1). They are then used to measure melodic similarity with respect to a target melody, whose definition is discussed in Section 2.2. Two different similarity measures are computed simultaneously: fundamental frequency (f_0) alignment (Section 2.3), and automatic singing transcription (Section 2.4) combined with note-level similarity (Section 2.5). The final step of the singing assessment system is the Performance rating stage (Section 2.6), which assigns an overall rating to the user performance.

2.1. Low-level feature extraction

We use the well-known Yin algorithm [11] to compute two related features: f_0 and aperiodicity (or voicing). These descriptors, combined with the instantaneous power of the audio

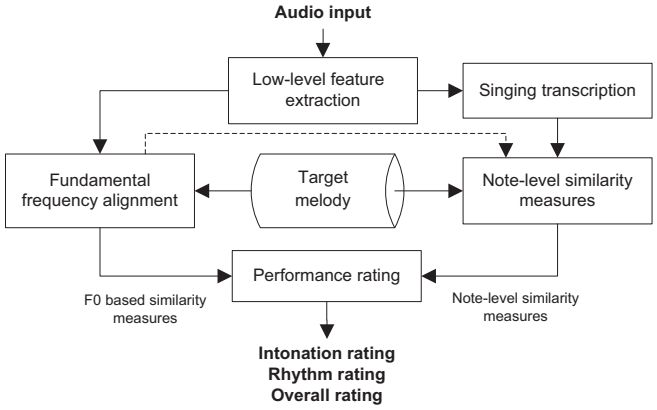

Fig. 1. Overall block diagram

signal, are given to other system blocks for singing assessment.

2.2. Target melody

The target melody is the performance that should be imitated by the student to achieve a good score. In our approach, the target melody is sung by is a *target singer*, i.e. a trained singer who is asked to sing with a rather pure voice, without vibrato, trying to be a good reference for beginners and children. Some post-processing is then applied to correct minor pitch and rhythm mistakes. Although we initially considered the symbolic score as a target melody, the fact of having a target singing voice allows a better alignment between f0 sequences and the measurement of detailed expressive resources.

2.3. Fundamental frequency alignment

Dynamic Time Warping (DTW) [12, 13, 14] is employed in order to find an optimal match between two given sequences under certain restrictions. However, it must be noted that the definition of optimal match strongly affects the robustness of the alignment. We have substituted the f_0 value of unvoiced regions by a constant value $f_{\text{unvoiced}} = 0$ Hz (see more details on voiced/unvoiced frame classification in Section 2.4). By removing the unvoiced sections, spurious f_0 values are avoided and only actual sung regions are compared. Therefore, the cost matrix M of the DTW can be defined as follow:

$$M_{ij} = min\{(f_{0T}(i) - f_{0U}(j))^2, \alpha\} \quad (1)$$

where $f_{0T}(i)$ is the f_0 value of the target melody in the frame i, $f_{0U}(j)$ represents the f_0 value of the user's performance in the frame j, M_{ij} is the cost value and α is a constant. When the squared f_0 difference becomes larger than α, it is assumed that an spurious case has been found and its contribution to the cost matrix is limited.

The DTW algorithm takes as input the cost matrix, and it provides an optimal path $[i_k, j_k]$ for $k \in 1 \dots K$, where K is the length of the path. We limit the slope of the path to the range $[10°, 80°]$ (deviations between transcription and reference are considered to be moderate).

2.3.1. DTW as an intonation similarity measure

The cost matrix provides information about the instantaneous deviation of the sung note with respect to the reference, as well as information about the total f_0 deviation of the sung melody. We consider the total cost of the optimal path to be a similarity measure for intonation assessment. The total intonation error (TIE) is computed as follow:

$$TIE = \sum_{k=1}^{K} M_{i_k j_k} \quad (2)$$

where M is the cost matrix, $[i_k, j_k]$ for $k \in 1 \dots K$ is the optimal path, and K is the length of the path.

2.3.2. DTW as a rhythmic similarity measure

In this paper, we propose DTW as a powerful procedure for automatic rhythm assessment. The idea is to analyze the shape of the optimal path, since it is a rich source of information about the rhythmic performance. In the cost matrix of the DTW, a $45°$ straight line represents a perfect rhythmic performance (no deviation with respect to the target melody). A poor rhythmic performance would yield deviations with respect to such straight line. The precise deviation location can be extracted from this curve, as well as the total amount of rhythmic error. On the other hand, a straight line with an angle $\alpha \neq 45°$ represents a good rhythmic performance in a different tempo. The straightness can be quantified through a linear regression analysis: Let $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{\varepsilon}$ be the linear model that best fits the optimal path within the cost matrix, with $\boldsymbol{\varepsilon}$ the error of fit. The error measure proposed is the root mean square (RMS): $\varepsilon_{\text{RMS}} = \sqrt{\frac{1}{N}\sum_{k=1}^{K} {\varepsilon_k}^2}$ in seconds. In Figure 2, two different situations are illustrated: a bad rhythmic performance that leads to a high linear regression error ($\varepsilon_{\text{RMS}} = 0.36s$, solid line), and the result of a perfect rhythmic performance played in a different tempo (dotted line). In the latter case, the linear regression error is very low ($\varepsilon_{\text{RMS}} = 0.047s$). Note that ε_{RMS} is a tempo-independent measure.

2.4. Singing transcription

We consider a f_0-based note segmentation approach with a hysteresis cycle for singing transcription [15, 16], and performed in the following steps: (1) locate the segments where the user is singing, (2) split the voiced segments into different notes and (3) label each note in terms of pitch.

We classify voiced and unvoiced frames by detecting stable frequency regions. If the f_0 curve is stable during a certain time (100ms in the implemented system), we create a new voiced segment. When there is a gap in the f_0 curve, such segment ends. Gaps of one exact octave are not considered,

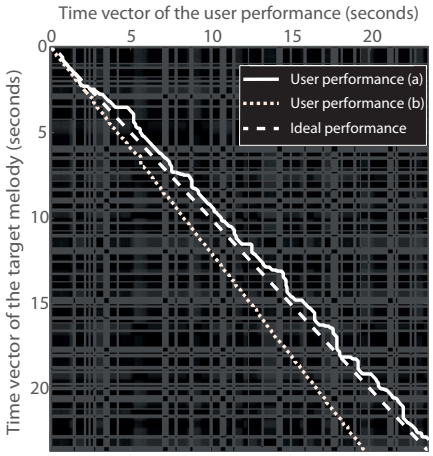

Fig. 2. Cost matrix of the DTW, together with the path for an ideal performance (dashed line) and two different user performances. Rhythmically unstable: $\varepsilon_{\text{RMS}} = 0.36s$ (solid line) and rhythmically stable (different tempo): $\varepsilon_{\text{RMS}} = 0.047s$ (dotted line).

since they are usually due to octave jumps during the same note. This process is carried out for the whole signal. In addition, voiced segments with a mean power below a threshold t_{pwr}, or mean aperiodicity above a threshold t_{ap} are directly tagged as unvoiced. This later classification avoids harmonic noises to be estimated as false voiced regions.

Once voiced segments are located, we segment voiced portions through f_0-based note segmentation. We use an hysteresis cycle in time and frequency in order to ignore minor deviations with respect to a pitch center. We dynamically estimate such pitch center by averaging f_0 values within a note. The estimated pitch average then becomes more precise as the note length increases. When the instantaneous f_0 of a note greatly deviates respect to its pitch average, a note split happens and the process starts again.

Once the sung notes are estimated, we assign a single pitch to each note. According to [17], the best pitch estimation for a note is a weighted mean of the most representative range of f_0 values. This type of mean is called alpha-trimmed mean [18], and it removes the extreme f_0 values (usually corresponding to the boundaries) before computing the mean. We have chosen this approach in this paper.

2.5. Note-level similarity measures

Note-level similarity measures are used to compare the symbolic representations of the sung melody and the target one. f_0 alignment, combined with melodic transcription, provides a note-to-note comparison even when rhythmic misalignment is present. The considered measures consist on the *average* ($\bar{x}$) and the *rhythmically weighted average* ($\bar{x}_W$) of three different magnitudes: onset time deviation (ΔO), note frequency deviation (Δf) and interval deviation (ΔI). While the *average* does not take into account note durations, the *rhythmically weighted average* does:

$$\overline{\Delta x} = \frac{\sum_{i=1}^{n} |\Delta x_i|}{n} \qquad \overline{\Delta x}_W = \frac{\sum_{i=1}^{n} l_i \cdot |\Delta x_i|}{\sum_{i=1}^{n} l_i} \quad (3)$$

where Δx_i is the deviation between the user transcription and target melody of the magnitude x (onset time, note frequency, or interval) for the note i, $\overline{\Delta x}$ is the average deviation of the generic magnitude x, $\overline{\Delta x}_W$ is the rhythmically weighted average deviation, l_i is the length of the note i, and n is the total number of notes. We now present the considered magnitudes.

2.5.1. Onset time

Let O_i be the onset time of the note i of the target melody, and $\hat{O}_i$ the onset time of the related note of the user performance. Then, the onset deviation is defined as $\Delta O_i = O_i - \hat{O}_i$.

2.5.2. Note frequency

We define f_i as the frequency of the note i of the target melody, and $\hat{f}_i$ as the frequency of the same note of the user performance. The note frequency deviation is then defined as $\Delta f_i = f_i - \hat{f}_i$ (where f_i is measured in cents in all cases) .

2.5.3. Interval

The interval is defined as the difference between the frequency of two consecutive notes $I_i = f_{i+1} - f_i$ in the target melody. The same interval in the user performance is defined as $\hat{I}_i = \hat{f}_{i+1} - \hat{f}_i$. The interval deviation is defined as $ID_i = I_i - \hat{I}_i$. This measure is key independent, so it is appropriated for a-cappella singing with no tuning reference.

2.6. Performance rating

In the performance rating stage, we combine the 8 similarity measures (2 DTW based and 6 at note-level) in order to provide three different ratings: rhythm rating, intonation rating, and overall rating. The optimal combination of the similarity measures has been considered to be the one that best fits the judgement of 4 trained musicians about 27 different singing performances. We have obtained such optimal combination through a quadratic polynomial regression analysis performed in Weka [19].

3. EVALUATION

3.1. Ground truth

We combine the use of real recordings and artificially generated melodies in order to systematically control the level of intonation and rhythm deviations. The evaluation dataset is then built by introducing random pitch/rhythm variations to three different target melodies, using an harmonic plus stochastic modelling of the input signal [20]. Three levels of random

variations have been applied for both pitch and rhythm. In total, nine combinations with different degree of error are generated from each reference melody. Therefore, 27 melodies (around 22 minutes of audio) comprise the whole evaluation dataset[1].

Human judgements were collected from four trained musicians, who were asked to score from 1 to 10 the evaluation dataset in three different aspects: intonation, rhythm and overall impression. Melodies were presented in random order using headphones.

3.2. Evaluation measures

Three different measures have been computed to evaluate the singing voice assessment system: interjudgement reliability, correlation between similarity measures and human judgements and polynomial regression error. Interjudgement reliability, proposed in [1], measures the correlation between human ratings. This measure aims to quantify the objectivity of the ratings. We have computed the correlation between the ratings for each pair of musicians (in total $n(n-1)/2 = 6$ pairs), and then averaged all the correlations. We have also computed the correlation coefficient for each similarity measure with respect to the different mean score given by musicians. This is a good reference about how meaningful each similarity measure is for performance assessment. A total of 27 (9 similarity measures $\times$ 3 ratings) correlation coefficients have been computed. Finally, the human criteria has been modelled in Weka through quadratic polynomial regression. The regression error quantifies the accuracy of the data fitting procedure. In this case, the evaluation dataset is the same as the training dataset. We consider the following measures from regression analysis: the correlation coefficient and the root mean squared error.

4. RESULTS & DISCUSSION

The mean correlation values corresponding to the interjudgement reliability measure are shown in Table 1. The results show that the agreement on rhythmic evaluation is lower. Nevertheless, the correlation in all cases is acceptable, and the case of intonation is specially good.

Type of score	Mean correlation coefficient
Intonation	0.93
Rhythm	0.82
Overall	0.90

Table 1. Results of interjudgement reliability

Table 2 shows the correlation between the different similarity measures and the human ratings. We observe a high correlation of human ratings and DTW based measures (TIE and ε_{RMS}), specially for rhythm assessment. DTW based measures do not require singing transcription, since it directly uses the low-level feature. Therefore, DTW is a simple but efficient technique for intonation and rhythm automatic assessment.

[1]Audio samples extracted from the ground truth can be found at http://www.atic.uma.es/singing

Similarity measure	Corr. with Intonation rating	Corr. with Rhythm rating	Corr. with Overall rating
TIE	0.92	0.21	0.81
ε_{RMS}	0.0012	0.81	0.52
$\overline{\Delta O}$	0.026	0.68	0.48
$\overline{\Delta O}_W$	0.037	0.68	0.48
$\overline{\Delta f}$	0.96	0.2	0.82
$\overline{\Delta f}_W$	0.89	0.23	0.82
$\overline{\Delta I}$	0.94	0.34	0.9
$\overline{\Delta I}_W$	0.87	0.35	0.87

Table 2. Correlation values of each similarity measure with the ratings given by trained musicians.

Type of error	Intonation	Rhythm	Overall
Correlation coefficient	0.988	0.969	0.976
Root mean squared error	0.4167	0.58	0.44

Table 3. Polynomial regression error.

Finally, Table 3 shows the obtained regression errors. The optimal polynomial combination of similarity measures provides high correlation with human judgements. For intonation, the results are specially good, because the chosen similarity measures are very representative and there is a high interjudgement reliability.

5. CONCLUSIONS

This paper presents a generic schema for automatic singing assessment, applied to the context of basic singing levels. The system provides the user with several ratings (intonation, rhythm and overall) by combining a set of melodic similarity measures with respect to a target melody. Target melodies are sung by a trained singer with neutral expression. The combination of f_0 alignment and symbolic similarity measures has been proven to be very appropriated for automatic rating. Furthermore, DTW based similarity measure is specially simple and effective for intonation and rhythm assessment, and such approach has not been considered in prior work. We have combined similarity measures through polynomial regression in order to fit the judgement of trained musicians. This approach then succeeds in modelling the musicians' criteria, as shown by our results. This study also contributes with a systematic evaluation methodology, applicable to other types of systems for automatic singing rating. Our approach is easily extensible to other expressive features such as vibrato or dynamics if new similarity measures are incorporated. In addition, the symbolic score of the melody could be used as target melody to avoid the need of a target singer. Finally, the proposed schema could be applied to realtime assessment if an on-line time warping algorithm [21] is integrated.

6. REFERENCES

[1] J. Wapnick and E. Ekholm, "Expert consensus in solo voice performance evaluation.," *Journal of voice official journal of the Voice Foundation*, vol. 11, no. 4, pp. 429–436, 1997.

[2] M. J. Bergee, "Faculty Interjudge Reliability of Music Performance Evaluation," *Journal of Research in Music Education*, vol. 51, no. 2, pp. 137, 2003.

[3] E. Ekholm, G. C. Papagiannis, and F. P. Chagnon, "Relating objective measurements to expert evaluation of voice quality in Western classical singing: critical perceptual parameters.," *Journal of voice official journal of the Voice Foundation*, vol. 12, no. 2, pp. 182–196, 1998.

[4] D. M. Howard, G. Welch, J. Brereton, E. Himonides, M. Decosta, J. Williams, and A. Howard, "WinSingad: a real-time display for the singing studio," *Logopedics Phoniatrics Vocology*, vol. 29, no. 3, pp. 135–144, 2004.

[5] Barcelona Music & Audio Technologies, "SKORE Performance Rating," *Internet, http://skore.bmat.me*, 2008.

[6] O. Mayor, J. Bonada, and A. Loscos, "The singing tutor: Expression categorization and segmentation of the singing voice," *Proceedings of the AES 121st Convention*, 2006.

[7] D. Rossiter and D. M. Howard, "ALBERT: a real-time visual feedback computer tool for professional vocal development.," *Journal of voice official journal of the Voice Foundation*, vol. 10, no. 4, pp. 321–336, 1996.

[8] Sony Computer Entertainment Europe, "Singstar," 2004.

[9] J. Callaghan and P. Wilson, *How to Sing and See: Singing Pedagogy in the Digital Era*, Cantare Systems, 2004.

[10] S. Grollmisch, E. Cano Cerón, and C. Dittmar, "Songs2see: Learn to play by playing," *Watermark*, vol. 1, 2012.

[11] A. De Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[12] Hiroaki Sakoe, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, pp. 43–49, 1978.

[13] C. A. Ratanamahatana and E. Keogh, "Everything you know about dynamic time warping is wrong," in *3rd Workshop on Mining Temporal and Sequential Data, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, 2004.

[14] D. Ellis, "Dynamic time warp (DTW) in Matlab," *Internet, http://labrosa.ee.columbia.edu/matlab/dtw. Last view: 29/11/2012*, 2003.

[15] I. Barbancho, C. de la Bandera, A.M. Barbancho, and L.J. Tardon, "Transcription and expressiveness detection system for violin music," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, april 2009, pp. 189 –192.

[16] E. Molina, E. Gómez, and Barbancho I., "Automatic scoring of singing voice based on melodic similarity measures," M.S. thesis, Universitat Pompeu Fabra, Music Technology Group, 2012.

[17] R. J. McNab, L. A. Smith, I. H. Witten, C. L. Henderson, and S. J. Cunningham, "Towards the digital music library: Tune retrieval from acoustic input," *Lloydia Cincinnati*, , no. 1978, pp. 11–18, 1996.

[18] J. Bednar and T. Watt, "Alpha-trimmed means and their relationship to median filters," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 32, no. 1, pp. 145–153, 1984.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[20] E. Gómez, G. Peterschmitt, X. Amatriain, and P. Herrera, "Content-based melodic transformations of audio material for a music processing application," *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx)*, 2003.

[21] S. Dixon, "Live tracking of musical performances using on-line time warping," *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx)*, 2005.