Automatic scoring of singing voice based on melodic similarity measures

Emilio Molina Martínez

MASTER THESIS UPF / 2012

Master in Sound and Music Computing

Master thesis supervisors:

Emilia Gómez

Department of Information and Communication Technologies

Universitat Pompeu Fabra

Isabel Barbancho

Departamento de Ingeniería de Comunicaciones

Universidad de Málaga

UNIVERSITAT POMPEU FABRA

Automatic scoring of singing voice based on melodic similarity measures

Emilio Molina

Music Technology Group
Universitat Pompeu Fabra
Tanger, 122-140, 3rd Floor
08018 Barcelona, SPAIN.

Master's thesis

Abstract A method for automatic assessment of singing voice is proposed. Such method quantifies in a meaningful way the similarity between the user performance and a reference melody. A set of melodic similarity measures comprising intonation and rhythmic aspects have been implemented for this goal. Such measure implement different MIR techniques, such as melodic transcription or score alignment. The reference melody is a professional performance of the melody, but the original score could be also used with minor changes in the schema. In a first approach, only intonation, rhythm and overall score have been considered. A polynomial combination of the similarity measures output are finally used to compute the final score. The optimal combination has been obtained by data fitting from a set of scores given by real musicians to different melodies. The teacher criteria is specially well modelled for pitch intonation evaluation. The general schema is also applicable to more complex aspects such as dynamics or expressiveness if some other meaningful similarity measures are included.

Computing Reviews (1998) Categories and Subject Descriptors:

H Information Systems
H.5 Information Interfaces and Presentation
H.5.5 Sound and Music Computing

Acknowledgements

I wish to thank my supervisors Emilia and Isabel for their guidance and advice. I also thank my family for their confidence in me, as well as my classmates for their support and for the good times we have experienced together. Special thanks to Xavier Serra for giving me the chance of being part of the SMC Master. Finally, I thank ATIC team of the University of Málaga for taking me into account for their research projects.

Contents

Chapter 1

Introduction

New information technologies have opened a wide range of possibilities for education. Nowadays, students can easily access to powerful resources than can be didactically exploited. Specifically, new portable devices such as smartphones, pads or laptops can be combined with complex signal processing techniques to enhance the capabilities of such didactic tools. On the other hand, current trends such as web 2.0 or cloud computing clearly set a framework that definitely is very interesting for educational purposes.

For the specific field of music, didactic applications usually take advantage of music information retrieval techniques. Such techniques can be efficiently implemented in different type of devices in order to provide a meaningful analysis of the student's performance.

This master thesis is framed in such context. It investigates about novel methods for an automatic assessment of music performances. Specifically, the adressed topic is the case of singing voice.

1.1 Motivation

Singing voice has been proved to be specially important during the music learning process. It strongly contributes to achieve a proper development of the musician skills (Welch et al., 1988).

The assessment of a singing performance is based on different criteria depending on the context and the age of the students. In the case of children and beginners, the evaluation criteria are mainly based on tuning, rhythm and the proper impost

of voice (in terms of energy and timbre) (Welch, 1994). Other advanced aspects such as vibrato or dynamics nuances are not taken into account at these levels.

Most of the existing systems are either oriented to entertainment, or they are designed as an auxiliary tool for a singing teacher (e.g. reviewed in Section 2.1.1). In general, they do not provide a tool for actual self-learning to the student. In this master thesis, novel techniques for automatic assessment of the singing performance are proposed. The novelty respect to previous system is an evaluation system based on a model of a real teacher to provide a helpful and complete feedback to the student.

1.2 Goals

The main goal is to develop novel methods for automatic assessment of the singing performance by modelling the criteria of a real teacher. The selected approach is based on melodic similarity measures of the user's performance respect to the reference melody. This goal is constraint to basic singing levels.

This aim is related to a sort of secondary goals:

1. Provide a state-of-the-art review in the fields of the music performance assessment, melody description and extraction and melodic similarity measures.
2. Elaborate an evaluation dataset
   (a) Recording of reference singing melodies. They can be post-processed with several software tools to correct any tuning or rhythm mistake.
   (b) Automatic processing of the signals in order to introduce controlled random variations of pitch and/or rhythm.
3. Develop a singing transcription algorithm: pitch estimation, note segmentation and parametrization.
4. Implement an score alignment algorithm.
5. Adapt the existing melodic similarity measures for the specific needs of the system.
6. Perform a regression analysis in order to model the criteria of real musicians.
7. Evaluate the system and discuss the results.

1.3 Structure of the thesis

1. Introduction: Motivation and goals of this master thesis.
2. State of the art: Relevant existing research about music performance assessment, melody description and extraction and melodic similarity measures.
3. Selected approach: Technical details about the selected approach for automatic singing assessment.
4. Evaluation methodology: Elaboration of the dataset and details about the evaluation measures.
5. Results and discussion: Obtained results and discussion about them.
6. Conclusions and Future work: Relevant conclusions and contributions, and some guidelines for future work.
7. References

Chapter 2

State-of-the-Art

In this literature review, current research about the main aspects of this master thesis will be analyzed and contextualized. Firstly, an overview on music performance evaluation will be presented. Some existing systems for automatic rating will be studied, as well as a musical perspective of the addressed problem. Most of the techniques to be implemented in this master thesis deal with such musical concepts. Then, the most relevant music information retrieval (MIR) techniques will be reviewed. These techniques will be organized into two sections: Melody description & extraction, and melodic similarity measures. Finally, in the last chapter some conclusions about the evaluation of the system have been extracted from previous research.

2.1 Music performance assessment

This master thesis aims to develop a system for the automatic rating of singing voice with pedagogic purposes. However, the scoring of a musical performance is not an easy task, even for expert musicians. In this section we present some previous approaches for automatic performance assessment, as well as a musicological study about the related problematic.

2.1.1 Existing systems for automatic evaluation

The systems for automatic rating of the singing voice have been typically applied in two fields: entertainment and educational applications.

Games and entertainment

In the last years, many musical games have been successfully commercialized. In the case of singing voice, the main approach is a karaoke-style game with automatic scoring. Some examples of these games are Singstar (Singstar, 2004), and other similar games (Ultrastar, Karaoke Revolution, etc.). These systems usually perform a rude analysis of the singing voice, and it usually takes into account just pitch and time.

Educational applications

The automatic assessment of singing voice with educational purposes typically lead to more complex systems. These systems should be able to provide a meaningful feedback to the user with the aim of improving the singing performance (like a virtual singing tutor). Songs2See is the most recent commercial system for this purpose, finally released in 2012 by Fraunhofer Institute (Dittmar et al., 2010). In (Mayor et al., 2006), a complete system for singing assessment based on pitch, rhythm and expressiveness accuracy is proposed. Such research finally lead to Skore (Skore, 2008), the system for online singer selection used in a famous reality TV show. Some other examples of previous educational systems are SINGAD (SINGing Assessment and Development)(Welch et al., 1988), WinSINGAD (Howard et al., 2004), ALBERT (Acoustic and Laryngeal Biofeedback Enhancement in Real Time) (Rossiter and Howard, 1996) and Sing & See (Sing&See, 2004). Some of the previous systems are rather oriented to provide low-level information about the singing voice, but they do not provide musical feedback for self-learning.

In general, all of them implement a meaningful performance analysis in real-time. However, the real-time approach can only give information about very short-time periods, and this doesn't model the complete judgment of an expert music teacher. Some other measures apart from real-time feedback are needed to really emulate the role of a music teacher. In the proposed system, this information for an appropriate assessment will be implemented by melodic similarity measures.

2.1.2 Musicological perspective

The assessment of a given musical performance is commonly affected by many subjective factors, even in the case of experts musicians' judgments. A sort of aspects such as the context, the evaluator's mood, or even the physical appearance of the performer (Griffiths and Davidson, 2006) can strongly change the perceived quality of the same performance. Thus, the development of an automatic performance

evaluation system seems to be a really challenging problem. However, under the correct conditions, some objectives aspects can be analyzed in order to model the expert's judgment.

Previous researchers have studied the reliability of judgments in music performance evaluation (Ekholm et al., 1998; Bergee, 2003; Wapnick and Ekholm, 1997), with some relevant results for the purposes of this master thesis. In such studies, different musicians were asked to grade a certain number of performers according to different aspects, with the aim to study how similar the different judgments were. In (Wapnick and Ekholm, 1997), the case of solo voice evaluation has been addressed. The different aspects to be evaluated in such experiment were rather technique: appropriate vibrato, color/warmth, diction, dynamic range, efficient breath management, evenness of registration, flexibility, freedom in vocal range, intensity, intonation accuracy, legato line, resonance/ring and overall score. Among these aspects, the ones presenting a higher reliability were intonation accuracy, appropriate vibrato, resonance/ring and the overall score. In the rest of experiments (Bergee, 2003), the rhythm/tempo aspects are also considered, and the conclusions are quite similar.

Such results are a good starting point in the automatic analysis of the performance. Since intonation, vibrato, timbre (resonances) and overall score seems to be more objectives aspects than the others (according to the reliability analysis), we will mainly focus our evaluation on these parameters. Rhythmic analysis will be also analyzed, since it can be easily evaluated for certain type of music material. In order to provide extra information for the overall score, an expressiveness evaluation of the performance will be also considered (phrasing, dynamics, etc).

2.2 Melody description and extraction

A good review about melody description and extraction techniques can be found in (Gómez et al., 2003). On the other hand, (Klapuri and Davy, 2006), presents some detailed information about melody transcription, with an specific approach for singing voice.

2.2.1 Pitch estimation

Pitch is the perceptual correlate of fundamental frequency, which is a physical measure. In this master thesis, we will use the term *pitch* referring to fundamental frequency, without perceptual considerations. In (Gómez et al., 2003), a general

review about the main methods for this purpose is presented. These techniques are classified in time-domain and frequency-domain approach. Two different techniques has been studied for the development of this master thesis:

- Yin algorithm (De Cheveigné and Kawahara, 2002): It is a time domain approach, and it can be considered as a improved version of the autocorrelation method.
- Two-Way Mismatch Method (Maher and Beauchamp, 1994): This is an harmonic matching method based on a frequency domain approach.

Other procedures such as zero-crossing rate estimation, or (Klapuri, 2003) approach have been discarded because they are either too simple or too complex. Between these two approaches, YIN algorithm has been the chosen technique for fundamental frequency extraction.

2.2.2 Note segmentation

The identification of notes from the original singing voice is a key task to achieve a good assessment of the performance. This is a problem very related to onset detection, since a note event can be identify from a similar approach. A good review on generic onset detection can be found in (Bello et al., 2005). However, the singing voice has some special features that lead to more specific algorithms.

An approach for note segmentation applied to singing voice is presented in (Viitaniemi et al., 2003) and (Ryyn et al., 2004). It describes note events with a hidden Markov model (HMM) using four musical features: pitch, voicing, accent and metrical accent. These features are used to estimate the transition between states of the note event: Attack, Sustain and Silence/Noise. In (Klapuri and Davy, 2006), this model is also exposed and detailed. This is the chosen approach in the system developed by (Mayor et al., 2006) for singing evaluation.

2.2.3 Extraction of note descriptors

Once the different notes have been segmented, a set of parameters have to be extracted from each one. In (Mayor, Oscar., Bonada, Jordi., Loscos, 2009), the considered parameters are pitch, volume, timing and expressive aspects such as vibrato or timbre.

According to (McNab et al., 1996) the perceived pitch of a note can be calculated by averaging the most representative pitch values into such time interval. This can be considered a mix between the mean and the mode of the pitch values. The whole energy is commonly computed with a simple average. Respect to the vibrato issue, in (Rossignol et al., 1999) a sort of procedures for its parametrization are reviewed.

2.2.4 Evaluation of the transcription accuracy

In (Ryyn et al., 2004), the transcription accuracy is evaluated by measuring the difference between a reference melody and the transcribed one. Two evaluation criteria were used: frame-based and note-based. The frame-based evaluation computed the error between the estimated pitch curve, and the reference. In the note-based evaluation, the hit ratio reflects the goodness of the system.

The case of melody extraction from polyphonic music is a more complex problem, and its evaluation usually takes into account more variables. The reviewed approaches for this type of evaluation also measure voicing and chroma accuracy (Poliner et al., 2007). The MIREX contest (MIREX, 2012) is also concerned about this problem, and similar evaluation procedures are proposed (MIREX, 2012). Despite singing transcription is a different problem, the related evaluation procedures can be useful to evaluate certain aspects of such task.

2.3 Melodic similarity measure

Melodic performance assessment, and melodic similarity are two related issues. A possible way to address the automatic assessment is by quantifying the similarity between the user performance and a target melody. This is the main idea behind the evaluation for the similarity measures proposed in (Mullensiefen and Frieler, 2004), and it is the selected approach in this master thesis.

Melodic similarity measures has been applied in many MIR tasks, such as query-by-humming systems (Pardo et al., 2004) or genre classification (Anan et al., 2011). A very interesting review on melodic similarity measures can be found in (Mullensiefen and Frieler, 2004). The same authors also implemented the toolkit SIMILE (Müllensiefen and Frieler, 2006). It consists on a set of implemented melodic similarity measures with a detailed documentation.

2.3.1 Musicological perspective

McAdams and Matzkin (2001) present a study on perceptual similarity from a musical point of view. They analyze the way we perceive similarity between two musical materials after applying a certain transformation. Such transformations are studied in different dimensions (mainly pitch and rhythm), and they evaluate the weight they affect the similarity perception and how are they interconnected. In such experiments, pitch and rhythm were initially considered as independent dimensions, and transformations were applied to each one in an independent way. However, the results showed a certain dependency between pitch and rhythm dimensions. Rhythmic variations in the same pitch pattern are usually perceived as more different than pitch variations in the same rhythmic pattern. On the other hand, a very important addressed point in (McAdams and Matzkin, 2001) is the importance of the musical "grammar". When studying grammatically coherent music (according to the tonal western style), transformations affecting the coherence of the music were perceived as more strong.

The results of the previous experiments lead to a sort of conclusions to be applied in this master thesis:

- The abstract information related to tonality and structure (in general "grammar" information) strongly affects the perception of similarity. Thus, these concepts should be somehow considered in a meaningful similarity measure.

- Overall similarity is perceived in different dimensions: pitch, duration, timbre, etc. According to the results of (McAdams and Matzkin, 2001), these dimensions are relatively independent, but not completely. The stored pitch information seems to be affected by rhythmic aspects, and that's also an important factor to be considered in the developed similarity measures. Rhythm can be taken as the skeleton of the music, that can really change the overall aspect of the above details (pitch, timbre, etc).

2.3.2 Representation of melodies and data transformations

Any measure of melodic similarity will necessarily be computed from a representation of the musical theme. The representation of the melody will affect the behavior of a given similarity measure, so it is an important aspect to take into account. In (Mullensiefen and Frieler, 2004), several melodic representations are proposed:

- [Duration, Pitch] series: Melody is represented as a series of a bidimensional points [Di,Pi]. Di makes reference to the inter-onset interval (IOI), and Pi to the absolute pitch position (MIDI note).
- [Duration, Interval] series: Instead of using the absolute pitch position, it uses the relative difference between consecutive pitches (intervals).
- Rhythmically weighted pitch series: In this case, the rhythmic information is stored in the number of times a certain pitch is repeated (e.g. [Di,Pi]=[1, 69], [2, 67] would be converted to wPi = [69, 67, 67].

The previous exposed melodic representations, ideally contain a complete description of the input melody. However, the simplification of the representations sometimes contributes to a similarity measure more related to the rough aspect of the whole melody.

2.3.3 Score alignment

When two melodies to be compared are rhythmically misaligned, a direct comparison over the pitch curve is meaningless. Due to that, the similarity measures should be complemented with a score alignment algorithm.

Cano et al. (1999) propose a method for score alignment of symbolic melodic representations based on hidden Markov models. However, it is not very appropriated for continuous curves. Other approaches are based on *Dynamic Time Warping* (DTW) for the alignment of two similar curves (Kaprykowsky and Rodet, 2006). This technique allows to find the optimal match between two vectors for aligning them. An implementation of a generic DTW algorithm can be found in (Ellis, 2003). This has been an important starting point in this master thesis. For possible real-time purposes, MATCH is a very interesting toolkit for dynamic score alignment (Dixon and Widmer, 2005).

2.3.4 Vector measures

If we consider the pitch series and the duration series as metrical vectors, we can perform some similarity measures by quantifying distances and projections between them. This kind of measures have been studied in (Aloupis et al., 2003), and they can found in the toolkit SIMILE (Müllensiefen and Frieler, 2006).

The proposed vector measures are the mean absolute difference (equation 2.1) and the correlation (equation 2.2)).

$$\mathrm{MAD}(x^1, x^2) = \frac{1}{N} \sum_{i=1}^{N} |x_i^1 - x_i^2| \tag{2.1}$$

$$\mathrm{corr}(x^1, x^2) = \frac{\sum_{i=1}^{N} x_i^1 \cdot x_i^2}{\sqrt{\sum_{i=1}^{N} x_i^1 \cdot x_i^1 \cdot \sum_{i=1}^{N} x_i^2 \cdot x_i^2}} \tag{2.2}$$

Where x_i^1 and x_i^2 are the two vectors of equal length N.

2.3.5 Musical measures

The use of the same scale into two different melodies can strongly affect to the perceived similarity between them. The predominant scale of a melody can be analyzed by a twelve-notes histogram, commonly called *chromagram*. The use of the chromagram vector for extracting tonal information from polyphonic audio data has been studied by (Gómez, 2006). The computation of the chromagram from symbolic information is even easier, since the histogram only takes into account the known pitch and duration of every note.

In (Mullensiefen and Frieler, 2004), two different types of harmonic similarity measures based on the Krumhansl-Schmuckler (Krumhansl, 1990) vectors are proposed:

- Harmonic vector correlation: For every bar of both melodies, the correlation with the Krumhansl-Schmuckler profiles are computed. The resulting vector-of-vectors from each melody are correlated bar by bar. Finally, the average correlation can be considered a harmonic similarity measures. Variations over this idea con provide some other harmonic vector correlations.

- Harmonic edit-distance: We compute a single tonality value for each bar as the key, which had the maximum value of the 24 possible keys, taking values 0-11 as major keys and values 12-23 as minor keys. This gave a "harmonic string" for each melody for which we can compute the edit-distance.

2.3.6 Evaluation of the similarity measures

We consider the methodology proposed by Mullensiefen and Frieler (2004), where the compared the results of such measures with an average of expert musicians

judgments. This is the chosen approach in order to evaluate further similarity measures.

MIREX contest is only oriented to symbolic similarity in the context of similar melodies retrieval, but it's an interesting evaluation procedure to take into account.

2.4 Evaluation

The evaluation of previous systems can be a good starting point to design a proper evaluation of the developed system. In the case of the singing scoring system presented in (Mayor, Oscar., Bonada, Jordi., Loscos, 2009), the evaluation has been performed with amateurs singers and five different pop songs. The accuracy in note segmentation, as well as expression regions, were evaluated to consider that the aim of the system was successfully achieved.

Other approaches have tried to study the influence of the system in a group of students during a certain time period. The evaluation of WinSingad (Howard et al., 2004) was performed in a singing studio with four adults students for an initial period of 2 months. A teacher was monitoring the evolution of the students, and his opinion was considered as a good feedback about the performance of the system.

A good evaluation should combine both approaches: the evaluation of the computational tasks comprising the system (such as transcription, similarity, etc.), as well as the representativeness of the final score for a musical self-learning of the student.

Chapter 3

Selected approach

The selected approach to perform an automatic assessment of singing voice is based on the schema shown in Figure 3.1.

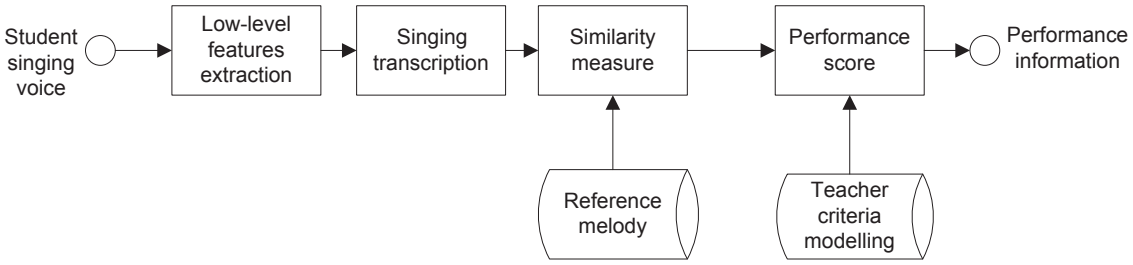

Figure 3.1: General schema of the proposed method for automatic singing assessment

3.1 Low-level features extraction

This block is based on Yin algorithm (De Cheveigné and Kawahara, 2002). This algorithm is based on the autocorrelation method, and it has become an standard for f_0 estimation in monophonic signals. Two meaningful signals are provided by this block: f_0 and aperiodicity (also called *degree of voicing*). These two curves, combined with the instantaneous power have been used to perform a note segmentation. The resulting curves have been smoothed with a median filter in order to avoid spurious change. Low-pass filtering has not been used because it affects to "sane" regions of the curves that could be helpful in later stages of the system.

3.2 Singing transcription

The selected approach for singing transcription is a pitch-based segmentation with a hysteresis cycle. This algorithm is one of the novelties of this master thesis, and it is an interesting approach for singing voice. In Figure 3.2, an example of melody has been transcribed to stable notes with the proposed algorithm.

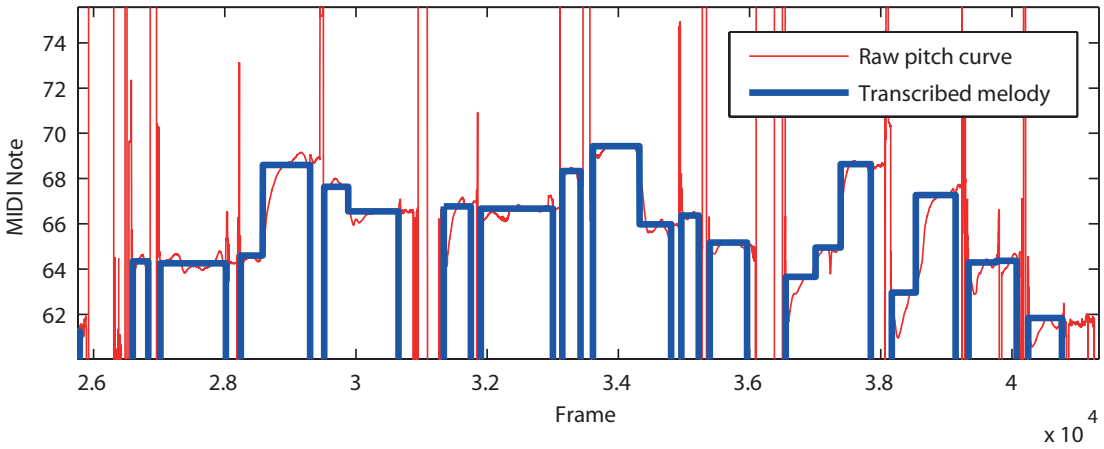

Figure 3.2: Original pitch curve and transcribed melody after applying the proposed singing method. Such method has been proved to be robust to instability of pitch.

The singing transcription block converts a monophonic input audio signal, to a symbolic music representation. This process allows to identify the different notes the user has sung for a later musical analysis. The singing transcription is performed in three steps:

1. Voiced / Unvoiced regions detection: It detects whether the user is singing or not. This process is commonly *voicing*, and it avoids spurious and/or not-detected notes.

2. Note segmentation: It splits the different sung notes within voiced segments.

3. Note pitch estimation: It assigns a constant pitch value to each estimated note.

3.2.1 Voiced/Unvoiced segments classification

The proposed approach is to detect stable frequency regions. If the f_0 is stable during 100 ms, a new segment starts and it is tagged with `f0_stable` = *true*. If a pitch gap is detected, the `f0_stable` flag is set to *false*. Gaps that are exactly one octave are not considered, since they are usually due to octave jumps during the same note. This process carries on until the whole signal has been processed.

Sometimes, unvoiced regions can present stable f_0 values if the environment noise is harmonic, or during certain fricatives consonants. Therefore, a more detailed classification is needed to properly decide among *voiced* and *unvoiced* segments. Three descriptors are computed for each segment:

- Duration of the longest region whose power is above a 20% of the mean power or all the previous segments: *longest_pwr_above20*
- Duration of the longest region whose aperiodicity value is below a threshold $t_{ap} = 0.18$: *longest_ap_below18*
- State of the `f0_stable` flag: *f0_stable*

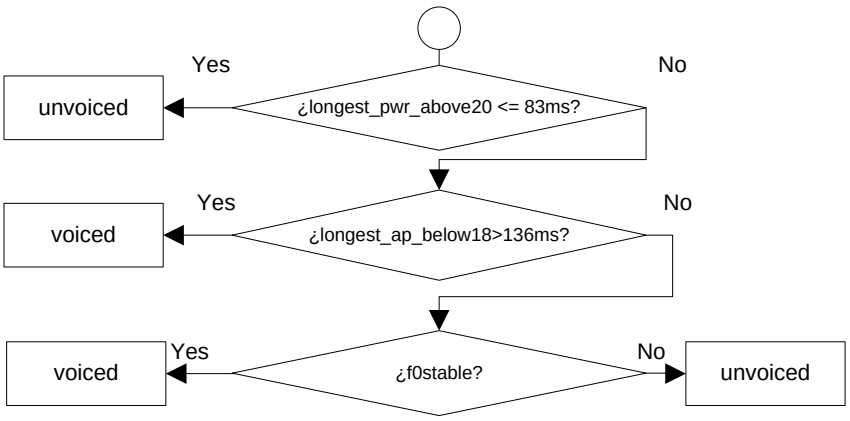

Figure 3.3: Implemented decision tree for voiced/unvoiced classification of segments.

A dataset of 2830 segments, manually labelled as *voiced* or *unvoiced* have been used to automatically generate a J48 decision tree classifier in Weka. The final used classifier is shown in Figure 3.3.

3.2.2 Pitch-based segmentation

Once the voiced regions are automatically identified, a second segmentation is needed to split legato notes. In the case beginner singers, the note segmentation becomes harder due the instability of pitch and energy within the same note. The proposed solution is a pitch-based segmentation with an hysteresis cycle in time and frequency. The hysteresis is a good approach to deal with unstable pitches. It is robust to minor variations, but it is sensitive to important and sustained changes in pitch. This method is partially based on (McNab et al., 1996) and (Ryyn et al., 2004).

This approach leads to the idea of *pitch centers*. When a note is been sung, minor deviations around a dynamically estimated pitch center are not considered. When a pitch deviation is sustained or very abrupt, it considers a note change and starts to compute a new pitch center. The estimation of such pitch center is performed by a dynamic averaging of the growing segment. Such average becomes more precise as the note length increases.

In Figure 3.4, the segmentation procedure is graphically shown. The left area between the actual pitch value and the average is measured at every frame. If such area overcomes a certain threshold, the note change happens and the whole process starts again.

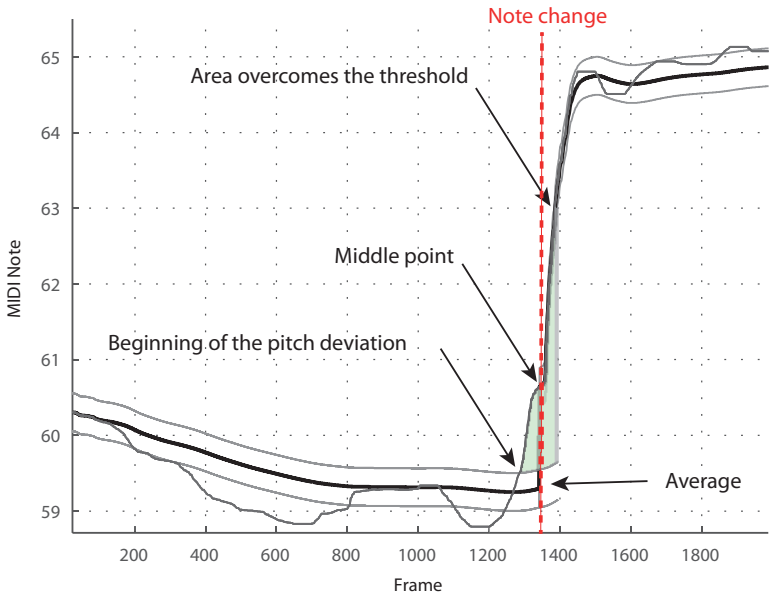

Figure 3.4: Graphical example of the proposed algorithm for note segmentation.

3.2.3 Note pitch estimation

Once the different sung notes have been segmented, a single pitch value has to be assigned to every note. According to McNab et al. (1996), the best pitch estimation for a note is a weighted mean of the most representative range of pitch values. This type of mean is called *alpha-trimmed mean*, and it removes the extreme pitch values (usually corresponding to boundaries) before computing the mean. In the chosen procedure, an energy weighted mean has been computed after discarding extreme pitch values.

3.3 Similarity measure

The automatic assessment of the singing performance is based on melodic similarity measures respect to a reference melody, considered as the "ideal" performance. In subsection 3.3.1, the chosen definition for reference melody is exposed.

When two melodies are rhythmically misaligned, a direct comparison between them can lead to meaningless results. Due to that, a score alignment based on dynamic time warping has been implemented (see subsection 3.3.2).

Next subsections presents the technical details about the developed similarity measures in this master thesis.

3.3.1 Reference melody

A key problem of musical performance assessment is defining the "ideal" performance, i.e. the *reference melody*. This reference melody can be defined in different ways, depending on the chosen assessment criteria. Two different approaches are interesting to define the reference melody:

- **Recording of a professional singer's performance:** In this case, the singer is asked to sing with a rather pure voice, without vibrato, and trying to be a good reference for beginners and children. Some post-processing with Melodyne (2010) has been applied to correct minor pitch. In such case, the professional musician agreed with the corrections.
- **Midi score:** On the other hand, the score of the melody can also be an interesting reference. However, it has not been used because score alignment did not offer such good results for specific cases. Further research should be needed for its robust implementation.

3.3.2 Score alignment

The selected approach for score alignment is based on *Dynamic Time Warping* (DTW). DTW is a method that allows a computer to find an optimal match between two given sequences under certain restrictions. However, the definition of *optimal match* strongly affects the robustness of the alignment. In this case, the alignment is optimized to fit the following conditions:

- The cost value to be minimized is the squared pitch difference between the user and the reference melodies. When two unvoiced frames are compared, the cost value is zero.
- A comparison between a voiced and an unvoiced frame should produce a controlled cost value.

This can be achieved by substituting pitch values of unvoiced regions by a very low constant value. On this way, meaningless pitch values are avoided. Then, the cost matrix M can be defined as follow: Let p_1 be the pitch series of the reference melody, and p_2 the pitch series of the user performance. The cost matrix is defined as: $M(i,j) = \min\{(p_1(i) - p_2(j))^2, \alpha\}$. When the squared pitch difference becomes higher than α, it is considered to be an spurious case and its contribution to the cost matrix is limited. It avoids that spurious pitch differences strongly affects the whole cost value.

The DTW algorithm takes as input the cost matrix, and it provides an optimal path $[i_k, j_k]$ for $k \in 1 \ldots K$, where K is the length of the path. Several restrictions are applied to avoid illogical situations, such as the alignment between two points that are too distant in time. More details about the DTW algorithm can be found in (Ellis, 2003). In Figure 3.5, an example of cost matrix together with the resulting time-warped pitch vectors are shown.

Score alignment as a similarity measure

Score alignment can be also considered a similarity measure. The shape of the path within the cost matrix gives an interesting measure about rhythmic deviations, whereas the accumulated cost-value of the path is a good reference about pitch accuracy. If the user performs with good rhythmic stability and exact tempo would produce a 45º line. On the other hand, good rhythmic stability but different tempo would produce straight lines with different angles. Curved lines represent instability and deviations respect to the original rhythm. Therefore, the *straightness* on the path is a good measure about the rhythmic performance.

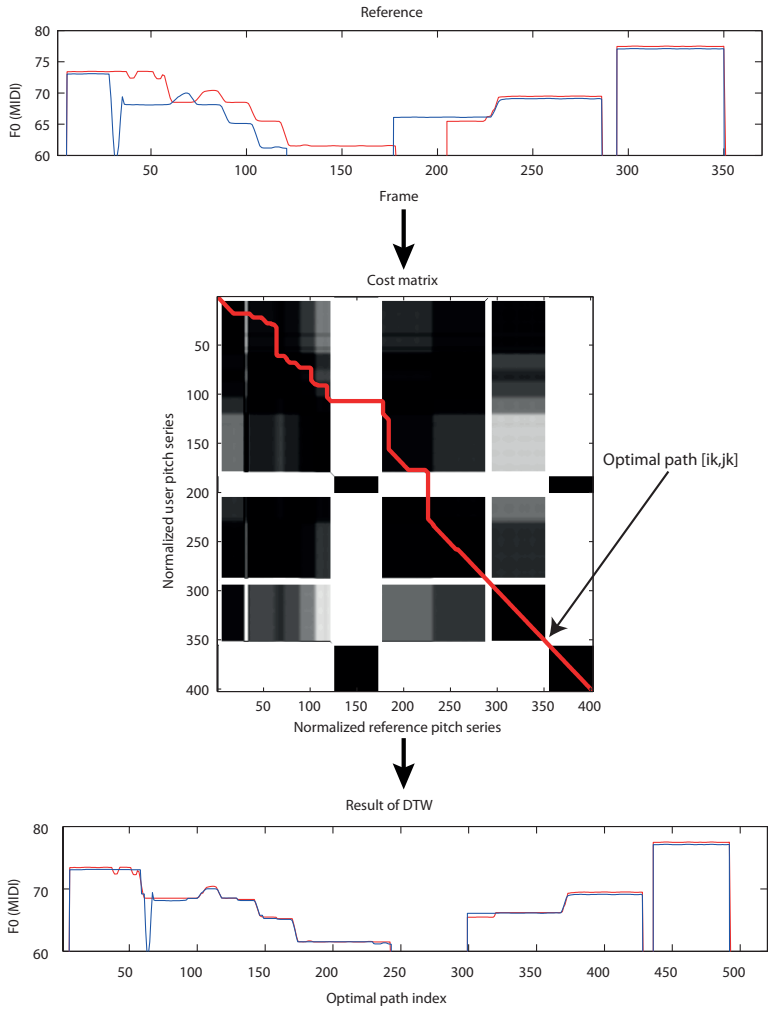

Figure 3.5: Dynamic time warping example. The red curve is the reference melody, and the blue one is the user's performance. After the optimal path, the alignment allows a proper comparison between melodies.

Rhythmic deviation: linear regression error

The straightness of the optimal path has been measure by performing a *linear regression*. The path values $[i_k, j_k]$ have been fitted into a polynomial of degree 1 by using the Matlab function `polyfit`. The mean squared difference between the original function and such polynomial is the linear regression error ϵ.

The linear regression error has been called: *lin_reg_err*.

3.3.3 Mean onset deviation

The combination of score alignment and note segmentation provides an interesting framework to perform different similarity measures. By combining these two techniques, the notes from the user performance can be directly associated to a note from the reference melody. Therefore, the same note can be identified in both melodies, even if they are not originally aligned in time.

The first interesting measure for rhythmic assessment is the mean onset deviation between notes. The problem of this measure is its low robustness against the onset imprecision during the note segmentation. For most of the sung melodies, the onsets should be precise enough to allow a meaningful similarity measure. The advantage of this measure is that it is quite close to the way musicians actually judge about rhythm.

The mean onset deviation has been called: *m_onset_dev*.

Rhythmically weighted mean onset deviation

This measure is a rhythmically weighted mean of the onset deviation. In this case, onsets belonging to long notes have a higher weight than short notes. The typical expression for a weighted mean is shown in (3.1).

$$\bar{x} = \frac{\sum_{i=1}^{n} \omega_i x_i}{\sum_{i=1}^{n} \omega_i} \tag{3.1}$$

Where $\bar{x}$ is the weighted mean, x_i is the signal and ω_i are the weights

The rhythmically weighted mean onset deviation has been called: *wm_onset_dev*
.

3.3.4 Mean pitch deviation

One of the most important aspects of singing assessment is the accuracy of intonation. The whole measure can be computed by measuring the mean absolute pitch deviation respect to the reference melody. This measure is **not** key independent, and just absolute pitch values are taken into account. Depending on the chosen criteria, this is not totally meaningful, because key is not critical for a-capella singing at basic levels.

The mean pitch deviation has been called: m_pitch_dev.

Since the previous measure does not take into account the duration of the notes, a rhythmically weighted mean is also proposed. In this similarity measure, long notes have a higher weight within the average.

The rhythmically weighted mean pitch deviation has been called: wm_pitch_dev.

3.3.5 Mean interval deviation

A way to normalize the key of the melodies is considering the interval deviation. The interval is defined as the pitch difference between two consecutive notes. In this case, the absolute key is not critical for the evaluation. This is a similarity measure more appropriated for a-capella singing.

The mean interval deviation has been called: m_interv_dev.

The rhythmically weighted version of this measure has been also included. It has been called: wm_interv_dev.

3.3.6 Harmonic profile correlation

According to Mullensiefen and Frieler (2004), the harmonic correlation is an interesting measure for melodic similarity, since it is representative of the whole sonority of a melody. In this case, the harmonic profile has been computed as an histogram of the importance for each pitch class within the melody. Such histogram is computed by summing the total duration of notes belonging to the same pitch class. The result is a chroma vector of 12 positions that contains interesting tonal information about the input melodic.

This is a key-dependent measure, and therefore it should be complemented with a key-independent measure.

The harmonic profile correlation has been called: h_prof_corr.

3.3.7 Interval profile correlation

The key-independent version of the previous measure is the interval profile correlation. In this case, a histogram of intervals belonging to the melody has been computed. This is representative of the whole sonority of the melody in a key independent way. For instance, a chromatic melody would strongly differ from a diatonic melody according to this measure.

The interval profile correlation has been called: $interv\_profile\_corr$.

3.4 Performance score

The final block of the singing assessment system if the *Performance Score*. It takes as input the similarity measure respect to the reference melody, and it gives a performance score to the user as a feedback to keep learning. In total, nine different similarity measures have been computed:

1. Linear regression error (rhythmic measure): lin_reg_err
2. Mean onset deviation (rhythmic measure): m_onset_dev
3. Rhythmically weighted mean onset deviation (rhythmic measure): wm_onset_dev
4. Mean pitch deviation (intonation measure): m_pitch_dev
5. Rhythmically weighted mean pitch deviation (intonation measure): wm_pitch_dev
6. Mean interval deviation (intonation measure): m_interv_dev
7. Rhythmically weighted mean interval deviation (intonation measure): wm_interv_dev
8. Harmonic profile correlation: $h\_profile\_corr$
9. Interval profile correlation: $interv\_profile\_corr$

These nine similarity measures is the input to the *Performance Score* block. The output consists on three different scores:

1. Intonation score
2. Rhythm score
3. Overall score

3.4.1 Teacher criteria modelling

The optimal combination of the nine similarity measures has been obtained by polynomial regression in Weka (Hall et al., 2009). The training dataset consists

on real scores given by trained musicians (at least 7 years of formal music studies) to a set of sung melodies. In total, 4 trained musicians have evaluated 27 different melodies, producing a training dataset of 108 instances for each score. This approach does not model a single teacher, but the average opinion of a group of teachers.

Chapter 4

Evaluation methodolody

The evaluation methodology is mainly based on two steps:

1. Dataset building: A dataset carefully designed has been built to perform a later evaluation of the performance of the system.
2. Computation of four evaluation measures:
   (a) Singing transcription accuracy: It measures the goodness of the singing transcription block.
   (b) Interjudgement reliability: It measures the correlation between the different opinions of the musicians.
   (c) Similarity measures correlation: It measures the correlation for each similarity measure with the scores given by the real musicians.
   (d) Polynomial regression error: It measures how well the system models the musicians judgement.

4.1 Dataset building

Due to the difficult of obtaining a big number of representative singing records, an alternative solution is proposed. The evaluation dataset has been generated by introducing random variations of pitch and rhythm to the reference melodies. Such melodic transformations are possible with an harmonic plus stochastic modelling of the input signal (Serra, 1989). For the case of singing voice, such model combined with the note segmentation definitely set an interesting framework to apply musical transformations.

Three different melodies of reference have been recorded. These melodies have been sung by a singing teacher, and post-processed with Melodyne to achieve a perfect rhythm and intonation. Three levels of random variations have been applied for both pitch and rhythm. In total, nine combinations with different degrees of mistakes are extracted from each reference melody. Therefore, 27 melodies (around 22 minutes of audio) comprise the whole evaluation dataset.

4.1.1 Harmonic plus stochastic model

In the proposed procedure, this model has been applied to every independent note. The typical steps to perform an harmonic plus stochastic modelling of the signal are:

1. Sinusoidal estimation
2. Harmonic matching
3. Extraction of the residual component
4. Stochastic modelling of the residual component

4.1.2 Random variations of pitch and rhythm

Pitch variations

The intervals of the melody have been modified in order to emulate the typical mistakes of beginners and children when they are singing. The whole contour of the melody is maintained, but the deviations of the intervals produce wrong pitch values. Three levels of interval modifications have been applied:

1. No variation: The pitch of the notes is not modified.
2. Weak interval variation: Every interval of the melody has been randomly modified. If the original interval is smaller than 4 semitones, a random pitch shifting between $[0, 0.8]$ semitones is applied. If the original interval is bigger than 4 semitones, such variation is comprised in $[0, 1.6]$ semitones. These values have been empirically chosen to achieve a realistic result.
3. Strong interval variation: For intervals smaller than 4 semitones, a random pitch shifting between $[0, 1.3]$ semitones is applied. If the original interval is bigger than 4 semitones, the variation is a random value between $[0, 2]$ semitones.

Rhythm variations

The same approach has been applied to the rhythmical transformations. Three levels of rhythmic variations have been considered:

1. No variation
2. Weak rhythmic variation: Each note has a random time stretching, whose ratio is comprised in $[60\%, 140\%]$.
3. Strong rhythmic variation: The ratio of the random time stretching is comprised in $[25\%, 170\%]$.

In real singers, the typical rhythmic mistakes are not independent for consecutive notes. Due to that, a slight low-pass filtering have been applied to the series of ratios in order to model the inertia of tempo variations.

4.2 Evaluation measures

Four different measures have been computed in order to evaluate the system. Such measures are presented at the beginning of this chapter, and they will be detailed in next subsections.

4.2.1 Singing transcription accuracy

The evaluation of the melodic transcription algorithm for singing voice is based on Ryyn et al. (2004) approach. Two different measures are computed:

- Note-based error: It does not take into account the duration, just the number of right notes.
- Frame-based error: It implicitly takes into account the duration of the notes, and it is more relevant for the needs of this master thesis.

These values are measured respect to manually annotated transcriptions. The annotations have been made in Cubase by a trained musician (10 years of music education) for 15 melodies randomly chosen from the dataset (around 12 minutes).

According to Ryynänen approach, the note-based evaluation is symmetrically approached from both the reference and the transcribed melodies' point of view. First, we count the number of reference notes that are *hit* by the transcribed melody and denote this number with $\check{c}_R$. A reference note is hit, if a note in the transcribed melody overlaps with the reference note both in time and in pitch. Second, the same scheme is applied so that the reference and transcribed melody exchange roles, i.e., we count the number of transcribed notes that are hit by the reference melody and denote the count with $\check{c}_T$. The *note error* E_n for a transcribed melody is the defined in (4.1).

$$E_n = \frac{1}{2}\left(\frac{c_R - \check{c}_R}{c_R} + \frac{c_T - \check{c}_T}{c_T}\right) \cdot 100\% \tag{4.1}$$

where c_R is the number of reference notes, and c_T is the number of transcribed notes.

The frame-based evaluation criterion is defined by the number of correctly transcribed frames c_{cor} and the number of voiced frames c_{ref} in the reference melody. A frame is considered to be correctly transcribed, if the transcribed note equals to the reference note in that frame. The *frame error* E_f for a transcribed melody is defined in (4.2).

$$E_f = \frac{c_{ref} - c_{cor}}{c_{ref}} \cdot 100\% \tag{4.2}$$

The frame and note errors are calculated for each individual melody in the evaluation database, and the average of these is reported.

4.2.2 Interjudgement reliability

The interjudgement reliability is an evaluation measured extracted from (Wapnick and Ekholm, 1997). It measures the correlation of the scores given by different musicians. This measure is useful to check the reliability and "objectivity" of the opinions. The correlation coefficient is a good way to check the coherence between two different musicians, and it can be computed as shown in (4.3).

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.3}$$

Where x_i are the scores given by one musician, and y_i are the scores given by another musician.

According to Wapnick and Ekholm (1997), in the case of having n musicians, a

good interjudgement reliability measure is the mean of the correlation coefficients for each pair of musicians. The total number of pairs for n musicians is $n(n-1)/2$. In this master thesis, 4 musicians have provide 3 different scores for 27 melodies. Therefore, the number of pairs analyzed is $4 \cdot 3/2 = 6$.

4.2.3 Similarity measures correlation

If a similarity measure is representative, the correlation with the musicians' scores should be high. The correlation coefficient has been computed for each similarity measure respect to the different mean scores given by real 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 scores) correlation coefficients will be computed.

4.2.4 Polynomial regression error

The teacher criteria modelling has been performed in Weka through polynomial regression. The regression error is the typical value for quantifying the accuracy of the data fitting. In this case, the evaluation dataset is the same as the training dataset. The provided measure about the regression analysis for a evaluation dataset x_i are:

- Correlation coefficient: see (4.3).
- Mean absolute error: $MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i|$, where $\hat{x}_i$ is the predicted value.
- Root mean squared error: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}$
- Relative absolute error: $RAE = \frac{\sum_{i=1}^{n}|x_i - \hat{x}_i|}{\sum_{i=1}^{n}|x_i - \bar{x}|}$ where $\bar{x}$ is the mean.
- Root relative squared error: $RRSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

Chapter 5

Results and discussion

In this chapter, the obtained results will be exposed and discussed. These results have been obtained with the selected approach and the previously exposed evaluation measures: Singing transcription accuracy, interjudgement reliability, similarity measures correlation and polynomial regression error.

5.1 Singing transcription accuracy

The obtained accuracy results for the proposed singing transcription system, according to Ryyn et al. (2004) evaluation measures are:

Note-based error: $E_n = 9\%$ (Ryynänen approach: $E_n = 9.4\%$)
Frame-based error: $E_f = 10\%$ (Ryynänen approach: $E_f = 9.2\%$)

This error is computed respect to set of manually annotated transcriptions. Despite the proposed singing transcription approach is simple, the obtained error is rather low, very close to state-of-the-art system such as Ryynänen approach. The typical errors are subsegmented notes, spurious notes and not detected notes. This kind of error, for the purpose of singing assessment are not critical. Therefore, the singing transcription algorithm is considered to be good enough for the scope of this master thesis.

5.2 Interjudgment reliability

Four trained musicians have been asked to score a set of 27 different melodies in three different aspects: intonation, rhythm and overall impression. However, the musicians' scores sometimes were not coherent. The reliability and the objectivity of the musicians for each aspect has been measured with the correlation coefficient.

For each pair of musicians ($n(n-1)/2 = 4 \cdot 3/2 = 6$ pairs), a correlation coefficient has been computed. The mean correlation values are shown in Table 5.1.

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

Table 5.1: Results of interjudgement reliability

The results show that agreement on rhythmic evaluation is more difficult. Nevertheless, the correlation in all cases is acceptable, and the case of pitch intonation is specially good.

5.3 Similarity measures correlation

Nine similarity measures have been computed. However, these measures are not equally meaningful for a later singing assessment. A good way to quantify the representativeness of each similarity measure, is by measuring the correlation with scores given by real musicians. If a high correlation between a similarity measure and the musicians' score is found, we will consider such measure as representative. This evaluation measure could be very useful for future improvements of the system, since meaningless similarity measures can be quickly detected.

In next subsections, the 27 correlation coefficients a graphically presented and organized according to the type of score: pitch intonation, rhythm and overall score.

5.3.1 Correlation with pitch intonation score

In Figure 5.1, the different similarity measures have been plotted respect to the musicians score for pitch intonation.

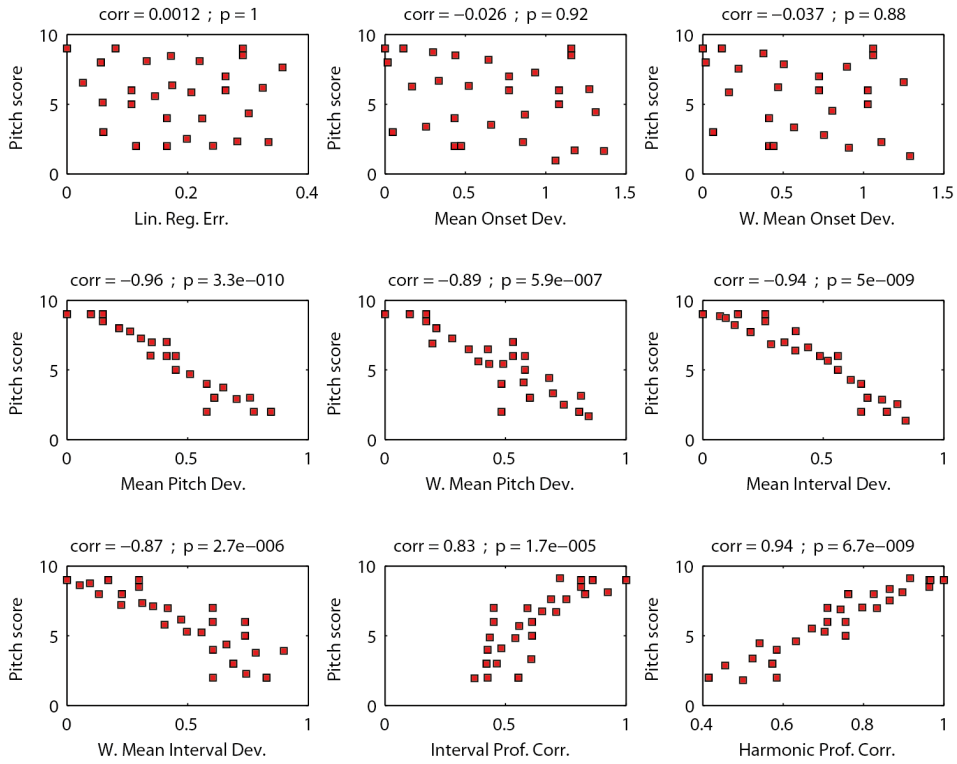

Figure 5.1: Output of each similarity measure vs. mean pitch intonation score given by real musicians. The correlation coefficient has been computed for each pair of magnitudes

Since some of the measures are computed for rhythm evaluation, they are not correlated with pitch intonation scores. However, some measures such as Mean Pitch Deviation, or the Mean Interval Deviation present a very interesting behavior. They are highly correlated with musicians scores, and therefore they are very representative for a pitch intonation evaluation of the singing performance.

5.3.2 Correlation with rhythm score

In Figure 5.2, the different similarity measures have been plotted respect to the musicians score for rhythm.

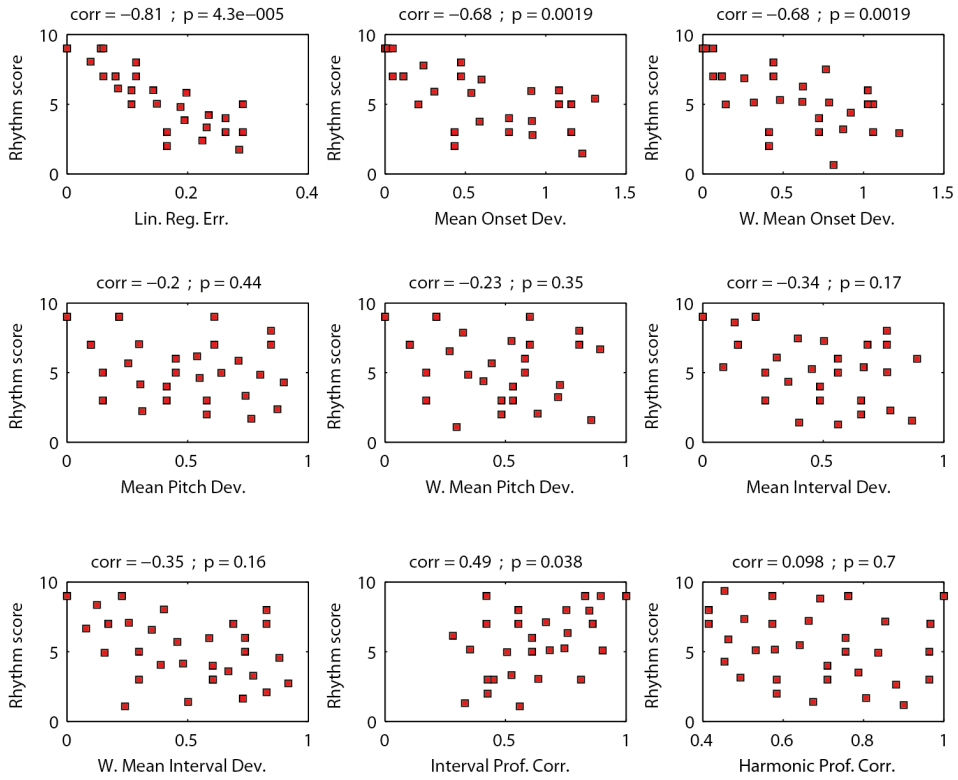

Figure 5.2: Output of each similarity measure vs. mean rhythm score given by real musicians.

In the case of rhythmic evaluation, just the *Linear Regression Error* is representative. It is surprising the low correlation of the measure called *Onset deviation*. This could be possible due to errors during the transcription, but a further analysis would be needed to really understand this lack of correlation. Another logical reason for this result could be the lower interjudgment reliability of the real musicians for rhythmic assessment.

5.3.3 Correlation with overall score

In Figure 5.3, the different similarity measures have been plotted respect to the musicians overall score.

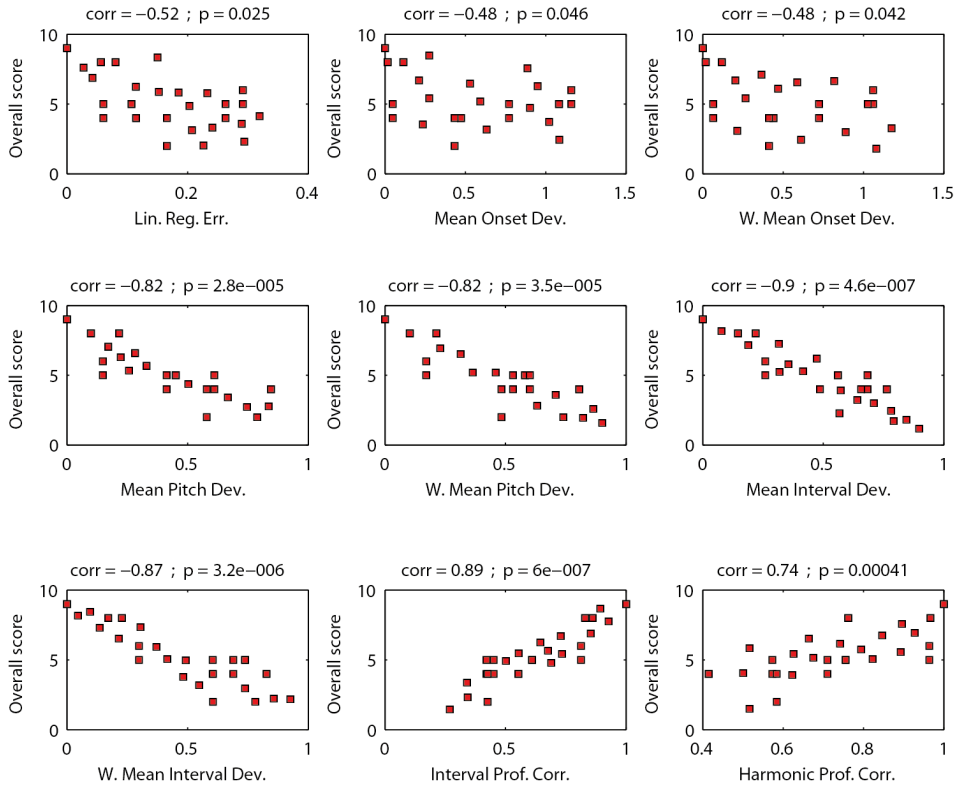

Figure 5.3: Output of each similarity measure vs. mean overall intonation score given by real musicians.

In this case, most of the computed similarity measures provide representative information. Specially, those measures related to pitch intonation provide information very correlated with the mean overall score. Therefore, musicians seem to give a higher weight to the pitch intonation in the overall impression.

5.3.4 General table of correlation coefficients

In Table 5.2, all the computed correlation coefficients are exposed. In such table, all the conclusions previously exposed can be observed. In addition, the *p-value* has been obtained for every correlation coefficient. Those coefficients with $p > 0.05$ are not representative, and they should not been taken into account.

Similarity measure	Correlation with intonation human score	Correlation with rhythm human score	Correlation with overall human score
lin_reg_err	C = 0.0012 (p = 1 > 0.05)	C = -0.81 ($p = 4 \cdot 10^{-5} < 0.05$)	C = -0.52 (p = 0.03 < 0.05)
m_onset_dev	C = -0.026 (p = 0.92 > 0.05)	C = -0.68 ($p = 2 \cdot 10^{-3} < 0.05$)	C = -0.48 (p = 0.046 < 0.05)
wm_onset_dev	C = -0.037 (p = 0.88 > 0.05)	C = -0.68 ($p = 2 \cdot 10^{-3} < 0.05$)	C = -0.48 (p = 0.042 < 0.05)
m_pitch_dev	C = -0.96 ($p = 3 \cdot 10^{-10} < 0.05$)	C = -0.2 (p = 0.44 > 0.05)	C = -0.82 ($p = 3 \cdot 10^{-5} < 0.05$)
wm_pitch_dev	C = -0.89 ($p = 6 \cdot 10^{-7} < 0.05$)	C = -0.23 (p = 0.35 > 0.05)	C = -0.82 ($p = 4 \cdot 10^{-5} < 0.05$)
m_interv_dev	C = -0.94 ($p = 5 \cdot 10^{-9} < 0.05$)	C = -0.34 (p = 0.17 > 0.05)	C = -0.9 ($p = 5 \cdot 10^{-7} < 0.05$)
wm_interv_dev	C = -0.87 ($p = 3 \cdot 10^{-6} < 0.05$)	C = -0.35 (p = 0.16 > 0.05)	C = -0.87 ($p = 3 \cdot 10^{-6} < 0.05$)
h_profile_corr	C = 0.83 ($p = 2 \cdot 10^{-5} < 0.05$)	C = 0.49 (p = 0.03 < 0.05)	C = -0.89 ($p = 6 \cdot 10^{-7} < 0.05$)
interv_profile_corr	C = 0.94 ($p = 7 \cdot 10^{-9} < 0.05$)	C = -0.098 (p = 0.7 > 0.05)	C = -0.74 ($p = 4 \cdot 10^{-4} < 0.05$)

Table 5.2: Correlation between similarity measures and musicians' judgements.

Those coefficients with $p > 0.05$ are not representative, and they should not been taken into account.

5.4 Polynomial regression error

Once the similarity measures have been computed, they have been combined in order to model the criteria of real musicians. This is a typical case of data fitting, and it has been addressed with polynomial regression in Weka. The final combination of values to fit each score has been shown in Figure 5.4. The different weights given to each similarity measure for each score are a good reference about its representativeness.

```
rhythm_score =

    -2.7651 * err_reglin +
    -7.8573 * onset_dev +
    -7.9661 * wonset_dev +
     0.5379 * pitch_dev +
     5.2152 * wpitch_dev +
     1.5955 * interval_dev +
     3.8211 * winterval_dev +
     4.9102 * int_prof_corr +
     0.2787 * harm_prof_corr +
     2.7752 * err_reglin^2 +
     6.0247 * onset_dev^2 +
     4.2181 * wonset_dev^2 +
    -1.0287 * pitch_dev^2 +
     4.3537 * wpitch_dev^2 +
    -3.7426 * interval_dev^2 +
    -1.4029 * winterval_dev^2 +
     1.1382 * int_prof_corr^2 +
     0.3818 * harm_prof_corr^2 +
     2.2911
```

```
pitch_score =

     2.2206 * err_reglin +
    -0.7465 * onset_dev +
    -0.6219 * wonset_dev +
    -0.8361 * pitch_dev +
     2.4439 * wpitch_dev +
    -1.2902 * interval_dev +
     1.8449 * winterval_dev +
     2.197  * int_prof_corr +
     1.2581 * harm_prof_corr +
    -0.4199 * err_reglin^2 +
     0.0841 * onset_dev^2 +
     0.0059 * wonset_dev^2 +
    -2.1229 * pitch_dev^2 +
     1.715  * wpitch_dev^2 +
    -4.7302 * interval_dev^2 +
    -0.672  * winterval_dev^2 +
    -0.2709 * int_prof_corr^2 +
     0.8888 * harm_prof_corr^2 +
     4.9271
```

```
overall_score =

    -0.2049 * err_reglin +
    -1.5564 * onset_dev +
    -1.5166 * wonset_dev +
    -0.6067 * pitch_dev +
     2.2424 * wpitch_dev +
    -1.1082 * interval_dev +
     1.2571 * winterval_dev +
     2.2069 * int_prof_corr +
     0.4329 * harm_prof_corr +
    -0.2842 * err_reglin^2 +
     0.5874 * onset_dev^2 +
     0.3847 * wonset_dev^2 +
    -1.3331 * pitch_dev^2 +
     2.0744 * wpitch_dev^2 +
    -3.2986 * interval_dev^2 +
    -0.1783 * winterval_dev^2 +
     0.6114 * int_prof_corr^2 +
     0.1724 * harm_prof_corr^2 +
     5.5764
```

Figure 5.4: Optimal combination of similarity measures to fit the musicians' judgement

The goodness of this regression determines the representativeness of the score given by the system. In Table 5.3, the different regression errors provided by Weka are shown.

Type of error	**Intonation**	**Rhythm**	**Overall**
Correlation coefficient	0.988	0.969	0.976
Mean absolute error	0.25	0.44	0.28
Root mean squared error	0.4167	0.58	0.44
Relative absolute error	10.345%	21.3%	15.67%
Root relative squared error	15.44%	24.36%	21.8%

Table 5.3: Polynomial regression error. This data is automatically provided by Weka.

The intonation score is the best result, because the chosen similarity measures are very representative and there is a high interjudgment reliability. The rhythm score is not so good, and deviation between predicted and estimated values are around 20%. Anyway, the results of rhythm score are interesting and it seems to be a good starting point for future improvements. The overall score is in the middle term, since it has a contribution from both pitch intonation and rhythmic aspects.

Chapter 6

Conclusions

In this master thesis, a method for automatic assessment of singing is proposed. This method is based on a similarity measure between the user's performance and a reference melody. Such reference melody is a recording from a singing teacher, as the ideal performance that the student should reach in terms of pitch and rhythm accuracy. The results of the different similarity measures have been combined in order to fit the judgement of real musicians. Such fitting has been performed through polynomial regression in the Weka environment (Hall et al., 2009). A novel singing transcription algorithm has been also implemented in order to allow note-to-note similarity measures. This algorithm is able to identify voiced segments and perform a pitch-based note segmentation. In addition, a score alignment algorithm has been included to properly compare both user and reference melodies when rhythmic deviations are present. It is based on *dynamic time warping* over the pitch curve (with some constraints). Such score alignment has been also used as a similarity measure, since rhythmic deviations can be directly extracted from it. The evaluation methodology is based on a dataset elaborated with randomly modified versions of the reference melodies. These manipulations have been performed with a pitch shifting / time stretching algorithm specially implemented for such purpose. The results after evaluation show that the chosen similarity measures are a good model of the criteria of real musicians, especially for the case of pitch intonation evaluation.

6.1 Contributions

According to the goals defined in Section 1.2, the contributions of the present study include:

- State-of-the-art in the most relevant fields for automatic performance assessment, with the spotlight on the case of singing voice.
- Generic system for automatic performance assessment. It is flexible and it can be easily extended to more complicated features, such as vibrato or dynamics.
- Novel algorithm for singing transcription based on a pitch-based note segmentation with a hysteresis cycle in pitch and time. It allows the note segmentation to be robust to unstable singing styles (appropriate for children and beginners).
- Set of similarity measures specifically designed for a later evaluation of the singing performance. They are complemented with a score alignment algorithm to deal with rhythmic misalignments.
- Dataset for singing evaluation based on random pitch shifting and time stretching over a reference melody.
- Algorithm based on the harmonic-plus-stochastic model for pitch shifting / time stretching of singing voice.
- Detailed evaluation of the system and discussion of the results.

6.2 Future work

In this section, the points in which the systems could be improved in near future are exposed.

- A real-time implementation of the system is proposed for future work. Visual feedback in real-time has been proven to improve the learning process for singing voice (Wilson et al., 2007). New causal similarity measures should be included for a real-time comparison between user's performance and reference melody. In such case, the final score would not be a number, but a curve along time.
- Rhythmic evaluation of the singing voice does not offer as good results as pitch intonation evaluation. A further research should be needed to find the real cause of this. Current similarity measures may be adjusted, and new rhythmic similarity measures could be included (currently there are only three).
- The dataset could be increased in number of reference melodies and musicians. This would serve to a better and more generic evaluation of the system.

- Some other aspects apart from pitch and rhythm could be included: vibrato, dynamics, etc. New descriptors and similarity measures should be implemented, but the schema of the developed system is generic enough to remain the same.

Bibliography

Aloupis, G., Fevens, T., Langerman, S., Matsui, T., Mesa, A., Nunez, Y., Rappaport, D., and Toussaint, G. (2003). Computing a Geometric Measure of the Similarity Between two Melodies. *Science*, pages 11–14.

Anan, Y., Hatano, K., Bannai, H., and Takeda, M. (2011). Music Genre Classification using Similarity Functions. *Learning*, 56(Ismir):693–698.

Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047.

Bergee, M. J. (2003). Faculty Interjudge Reliability of Music Performance Evaluation. *Journal of Research in Music Education*, 51(2):137.

Cano, P., Loscos, A., and Bonada, J. (1999). Score-Performance Matching using HMMs. In *Proceedings of the International Computer Music Conference ICMC*, volume 1, pages 441–444. Citeseer.

De Cheveigné, A. and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917.

Dittmar, C., Großmann, H., Cano, E., Grollmisch, S., Lukashevich, H. M., and Abeßer, J. (2010). Songs2see and globalmusic2one: Two applied research projects in music information retrieval at fraunhofer idmt. In Ystad, S., Aramaki, M., Kronland-Martinet, R., and Jensen, K., editors, *CMMR*, volume 6684 of *Lecture Notes in Computer Science*, pages 259–272. Springer.

Dixon, S. and Widmer, G. (2005). MATCH: A Music Alignment Tool Chest. In Reiss, J. D. and Wiggins, G. A., editors, *Proc ISMIR London GB*, number Ismir, pages 492–497. University of London.

Ekholm, E., Papagiannis, G. C., and Chagnon, F. P. (1998). 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*, 12(2):182–196.

Ellis, D. (2003). Dynamic Time Warp (DTW) in Matlab. Web resource available: http://www.ee.columbia.edu/ dpwe/resources/matlab/dtw/.

Gómez, E. (2006). *Tonal Description of Music Audio Signals*. PhD thesis, University Pompeu Fabra, Barcelona, Spain.

Gómez, E., Klapuri, A., and Meudic, B. (2003). Melody Description and Extraction in the Context of Music Content Processing. *Journal of New Music Research*, 32(1):23–40.

Griffiths, N. and Davidson, J. (2006). The effects of concert dress and physical appearance on perceptions of female solo performers. In *9th International Conference on Music Perception and Cognition*, pages 1723–1726.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Howard, D. M., Welch, G., Brereton, J., Himonides, E., Decosta, M., Williams, J., and Howard, A. (2004). WinSingad: a real-time display for the singing studio. *Logopedics Phoniatrics Vocology*, 29(3):135–144.

Kaprykowsky, H. and Rodet, X. (2006). Globally Optimal Short-Time Dynamic Time Warping, Application to Score to Audio Alignment. *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 5:249–252.

Klapuri, A. (2003). Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, pages 804–816.

Klapuri, A. and Davy, M. (2006). *Signal Processing Methods for Music Transcription*. Springer.

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch*, volume 17. Oxford University Press.

Maher, R. C. and Beauchamp, J. W. (1994). Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *Journal of the Acoustical Society of America*, 95(4):2254–2263.

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

Mayor, Oscar., Bonada, Jordi., Loscos, A. (2009). Performance analysis and scoring of the singing voice. *Proc 35th AES Intl Conf London UK*, pages 1–7.

McAdams, S. and Matzkin, D. (2001). Similarity, invariance, and musical variation. *Annals Of The New York Academy Of Sciences*, 930(1):62–76.

McNab, R. J., Smith, L. A., and Witten, I. H. (1996). Signal Processing for Melody Transcription. *Proceedings of the 19th Australasian Computer Science Conference*, 18(4):301–307.

Melodyne (2010). Melodyne Editor by Celemony. http://www.celemony.com.

MIREX (2012). Music Information Retrieval Evaluation eXchange contest, retrieved from www.music-ir.org/mirex.

Mullensiefen, D. and Frieler, K. (2004). Cognitive Adequacy in the Measurement of Melodic Similarity: Algorithmic vs. Human Judgments. *Computing in Musicology*, 13(2003):147–176.

Müllensiefen, D. and Frieler, K. (2006). The SIMILE algorithms documentation.

Pardo, B., Shifrin, J., and Birmingham, W. (2004). Name that tune: A pilot study in finding a melody from a sung query. *Journal of the American Society for Information Science and Technology*, 55(4):283–300.

Poliner, G. E., Ellis, D. P. W., Ehmann, A. F., Gomez, E., Streich, S., and Ong, B. (2007). Melody Transcription From Music Audio: Approaches and Evaluation. *Ieee Transactions On Audio Speech And Language Processing*, 15(4):1247–1256.

Rossignol, S., Depalle, P., Soumagne, J., Rodet, X., and Collette, J.-L. (1999). Vibrato: detection, estimation, extraction, modification. In *Notes*, volume 99, pages 3–6. Citeseer.

Rossiter, D. and Howard, D. M. (1996). ALBERT: a real-time visual feedback computer tool for professional vocal development. *Journal of voice official journal of the Voice Foundation*, 10(4):321–336.

Ryyn, M. P., Klapuri, A. P., Box, P. O., and Tampere, F. (2004). Modelling of Note Events for Singing Transcription. In *Signal Processing*. ISCA, Citeseer.

Serra, X. (1989). *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University.

Sing&See (2004). Sing&See by CantOvation Ltd. Website: http://www.singandsee.com.

Singstar (2004). Singstar game, by Sony Computer Entertainment Europe. http://www.singstar.com/.

Skore (2008). Skore by Barcelona Music&Audio Technologies (BMAT). Retrieved from http://www.bmat.com/cas/productos/skore/index.php.

Viitaniemi, T., Klapuri, A., and Eronen, A. (2003). A probabilistic model for the transcription of single-voice melodies. In *Proceedings of the 2003 Finnish Signal Processing Symposium FINSIG'03*, number 20, pages 59–63. Tampere University of Technology, Citeseer.

Wapnick, J. and Ekholm, E. (1997). Expert consensus in solo voice performance evaluation. *Journal of voice official journal of the Voice Foundation*, 11(4):429–436.

Welch, G. F. (1994). The assessment of singing. *Psychology of Music*, 22(1):3–19.

Welch, G. F., Rush, C., and Howard, D. M. (1988). The SINGAD (SINGing Assessment and Development) system: First applications in the classroom. *Proceedings of the Institute of Acoustics*, 10(2):179–185.

Wilson, P. H., Lee, K., Callaghan, J., and Thorpe, C. W. (2007). Learning to sing in tune: Does real-time visual feedback help? *CIM07 3rd Conference on Interdisciplinary Musicology Tallinn Estonia*, 2(1):15–19.