Addressing user satisfaction in melody extraction

Belén Nieto

MASTER THESIS UPF / 2014

Master in Sound and Music Computing

Master thesis supervisors:

Emilia Gómez

Julián Urbano

Justin Salamon

Department of Information and Communication Technologies

Universitat Pompeu Fabra, Barcelona

Acknowledgment

I want to thank everyone who has made possible this research, specially to the people responsible for the SMC master admissions for making my dream come true, my supervisors for their support, guidance and for teaching me the importance and value of the evaluation tasks, my fellows for being so nice and all the moments we spent together, the people who kindly participated in the survey for their time and of course the authors who provided the outputs of their algorithms. I also would like to thank and congratulate all the MTG for the wonderful research atmosphere they create. Moreover, I want to express my appreciation to the UPF for the financial support. Last but not least, thanks to my family for their unconditional support.

Heartfelt thanks to all for this amazing and unforgettable experience.

Abstract

The aim of this master thesis is to carry out a perceptual evaluation of the melody extraction task. To do so, it has been conducted a survey in which 26 subjects have participated in order to test some hypotheses about how the different errors affect the perception of the melodies by the users. In view of the results it can be checked how different kind of errors have a different impact in the quality perceived by the users, in such a way that the result of the perceptual evaluation need not to be equal to the one achieved by the current evaluation metrics. Finally, this research shows that there is much scope for future work refining new evaluation metrics to better capture user preferences.

Resumen

El objetivo de este Trabajo Fin de Máster es llevar a cabo una evaluación perceptual de la tarea de extracción de melodías. Para ello, se ha realizado una encuesta con un total de 26 participantes para comprobar algunas hipótesis acerca de cómo los diferentes errores afectan la percepción de las melodías por los usuarios. A la vista de los resultados se puede comprobar como diferentes tipos de errores tienen un impacto diferente en la calidad percibida por los usuarios, de tal manera que el resultado de la evaluación perceptual no tiene por qué coincidir necesariamente con el resultado de la evaluación utilizando las métricas actuales. Finalmente, esta investigación pone de manifiesto que aún queda mucho por hacer en el futuro refinando nuevas métricas de evaluación para capturar mejor las preferencias de los usuarios.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

In this chapter, first we explain the motivation of our research and our aims. Then, we present our methodology and the means that we are going to use. Finally, we expose the outline of this document.

1.1 Motivation and aims

Until now most of the evaluations of Music Information Retrieval (MIR) tasks are based only on system-centered metrics, being the ones used at MIREX (Music Information Retrieval Evaluation eXchange) the *de facto* standard. However, in most of the cases the opinion of the users is disregarded due to the difficulty of obtaining their judgements, because carrying out a survey involving people may result expensive compared to the ease of using only system-centered metrics. That is the case of the melody extraction task.

So, the main motivation of this master thesis is the perceptual evaluation of the melody extraction task. Moreover, we will also check if the current metrics reflect properly the opinion of the users. Furthermore, if they do not represent the perception of the users, which is the expected case, we will also try to find a new metric similar to the current ones but more representative of user preferences.

1.2 Means and methodology

The means employed during the developing of this master thesis are the following:

- Sonic Visualizer for the preliminary analysis of the different errors.
- Matlab.
- Website based on HTML, PHP and SQL.
- The technology provided by Hostinger and Google Sites.

- Dataset containing 30 excerpts, their ground truth annotations and the output of six different melody extraction algorithms: Dressler (2009), Durrieu et al. (2010), Fuentes et al. (2012), Hsu and Jang (2010), Paiva et al. (2006) and Salamon and Gómez (2012).

The methodology employed to reach the aims of this master thesis is the following:

- First, the outputs of the different algorithms and the ground truth annotations are synthesized and carefully listened and analysed, paying special attention to the different kind of errors and their distribution.
- Then, accordingly to our perception we propose some hypotheses about how the different errors affect the perceived quality of the excerpts.
- Next, we have to design an experiment to test the different hypotheses by means of a survey involving the opinion of the users.
- Finally, after analysing the results of the survey we should be able to propose a new metric that represents better the perception of the users than the current ones.

1.3 Outline

The outline of the document is the following:

1. **Introduction**: This chapter presents the motivation and the methodology of our research.
2. **State-of-the-art**: This chapter exposes the current state of the melody extraction algorithms and the user-centered evaluation approaches found in the literature.
3. **User survey**: In this chapter it is presented the designing of the experiment involving the users and the results of the survey.
4. **Conclusion**: Finally, it is explained the conclusion and the future perspectives.

Chapter 2

STATE OF THE ART

This chapter overviews the current state of the art in melody extraction. In order to do that, first the framework of the task is defined. Then, a classification of the existent algorithms is proposed. Moreover, some relevant algorithms for the development of this master thesis are reviewed. Next, the current procedure for evaluating the task is analysed. Finally, we discuss about the relevance of user-centered approaches for the evaluation of MIR tasks. To conclude, we review a set of MIR algorithms involving the user.

2.1 Melody extraction

This section presents the framework of the melody extraction task, the classification and review of the different algorithms proposed in the literature (Salamon, 2012).

2.1.1 Definition of the task

The aim of melody extraction is to automatically obtain the melody pitch contour of the predominant instrument in an audio signal of polyphonic music (Fig. 2.1). Because of the controversy about some of the terms included in the previous statement, it is necessary to clarify some concepts in order to define the framework of the task (Salamon et al., 2014) :

- **Musicological concept of "melody":** The concept of melody has been discussed for a long time, but it has not been achieved yet an agreement on its definition. Moreover, it is a notion that ultimately depends on the subjective perception of human listeners and therefore, its meaning will be context dependant. However, in order to share the same interpretation the MIR community has adopted the following definition, proposed by Poliner et al. (2007): *"The melody is the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music, and that a listener would recognize as being the 'essence' of that music when heard in comparison"*.

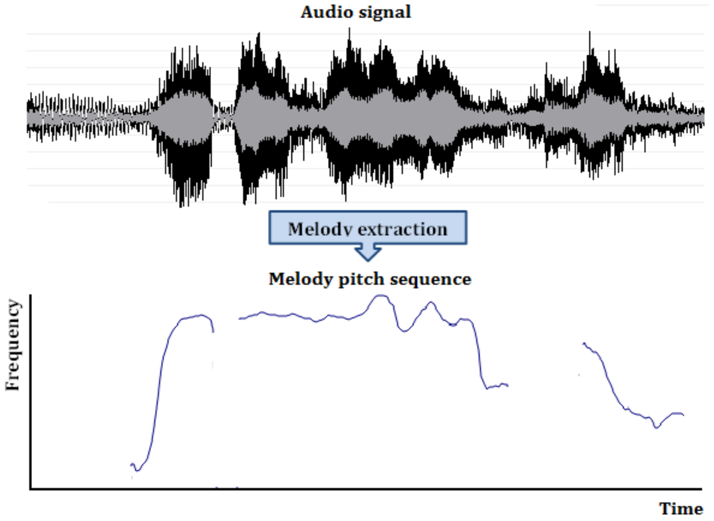


Figure 2.1: Illustration of the aim of melody extraction: obtaining a sequence of fundamental frequency values representing the predominant melody of an audio signal.

- **Concept of predominant instrument:** Given the previous definition of melody, the task is still subjective because each listener may select a different melodic line (e.g., one listener may focus on the lead vocals while other prefers the solo played by another instrument). In order to limit the range of possible answers, the task is limited to the estimation of the melody played by a single sound source which is considered the most predominant in the mixture. Although this decision is still debatable, because not everybody would agree in what predominant means, this problem is overcome in the practice by working with music collections with a clear leading instrument.

- **Concept of melody pitch contour**: First, the pitch is a perceptual notion which was defined by ANSI (American National Standards Institute) as an attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high. It mainly relies on the frequency content of the sound, but it also depends on the sound pressure and the waveform (Hartmann, 1997). Secondly, the fundamental frequency is a physical quantity which corresponds to the lowest frequency of a periodic waveform. However, in the literature related to melody extraction it is common to use the term pitch to refer to the fundamental frequency. So, the pitch contour in this context is a sequence of fundamental frequency values that jointly represent the melody.

- **Concept of polyphonic**: Finally, in this context the term polyphonic refers to music in which two or more notes can sound simultaneously. But this does not necessarily imply the presence of various instruments (e.g. voice and guitar), because the notes could also have been played by a single instrument capable of playing various notes at the same time (e.g. piano).

To conclude, taking into account the previous considerations, it is important to note that the melody extraction task consists mainly of two different parts:

- **Voicing detection**: In this part the algorithm must decide for a given time instant if the predominant melody is present or not. That is, if the main instrument is playing or not at that moment.
- **Pitch detection**: In this step the algorithm must decide the most likely fundamental frequency of the note played by the predominant instrument for each time instant.

2.1.2 Melody extraction algorithms

This section presents a classification of the melody extraction algorithms. Then, the relevant algorithms to this master thesis are reviewed.

2.1.2.1 Classification

According to Salamon et al. (2014) the melody extraction algorithms can be classified into three groups: salience based-approaches, source-separation approaches and alternative approaches. The main characteristics of each group are reviewed below.

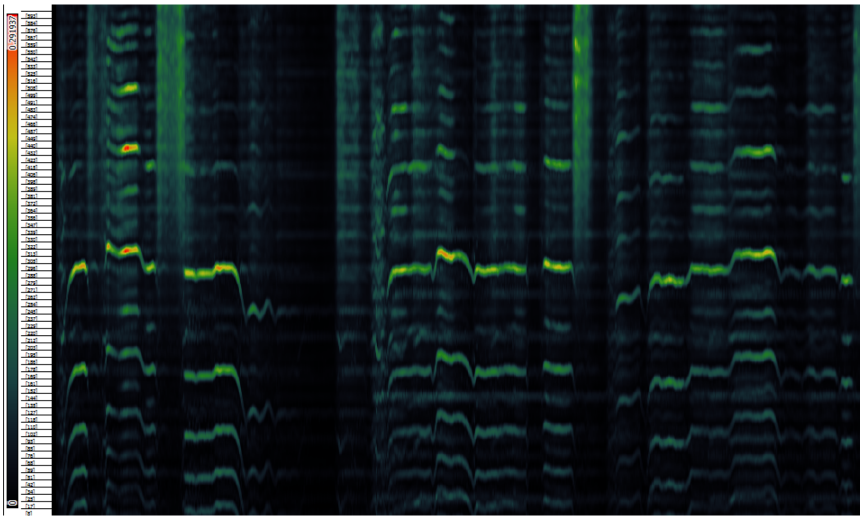


Figure 2.2: Example of the salience function of a pop excerpt containing only singing voice and guitar, computed using the MELODIA vamp-plugin proposed by Salamon and Gómez (2012). The horizontal dimension represents the time, and the vertical dimension denotes the pitch candidates.

Salience based-approaches: This group includes all the algorithms based on the construction of a pitch salience function, which is a time-frequency representation of pitch salience calculated from the mixed signal. There are various approaches for computing this function. For instance, a simple solution would be to compute the salience of each candidate frequency as the weighted sum of its harmonics (Fig. 2.2). The candidate frequencies are those included in the range in which it is expected to find the predominant melody. Such algorithms typically involve seven steps (Fig.2.3):

1. *Preprocessing*: First, some approaches include a preprocessing applied to the audio signal. The most common options are the following:

    - Filtering in order to enhance the frequency bands in which it is expected to find the predominant melody. For example, Salamon and Gómez (2012) employ an equal loudness filter, while Goto (2004) uses a band pass filter between 261.6Hz and 4KHz.
    - Using source separation in order to enhance the predominant melody signal. For instance, Hsu and Jang (2010) employ an adaptation of harmonic-percussive sound separation (HPSS) in order to separate the melody from the accompaniment.

2. *Spectral transform*: Next, the audio signal is divided into frames and a transform function is applied in order to obtain a spectral representation of each frame. Some approaches employed at this step are the following:

    - Short-Time Fourier Transform (STFT): This is the simplest approach and the one employed by most of the proposed algorithms (Ryynänen and Klapuri, 2008). Typically it is used a window size between 50 and 100 ms, which usually provides enough frequency resolution to distinguish between close notes while maintaining a sufficient time resolution to appreciate fast pitch changes in the melody.
    - Multiresolution transforms: These approaches try to solve the time-frequency resolution limitation due to the Fourier Transform by using larger windows at low frequencies and smaller windows at high frequencies. Some examples of this techniques are the multirate filterbank (Goto, 2004), the constant-Q transform (Cancela, 2008) or the Multi-Resolution FFT (MRFFT) (Dressler, 2006).
    - Human auditory system: Finally, some algorithms use transforms designed especially to emulate the human auditory system (Paiva et al., 2006).

3. *Computation of the spectral peaks*: Once the transform has been applied, most approaches only use the peaks of the spectrum for further processing. So, the next step is to compute these peaks. Then, some algorithms apply peak processing techniques such as:

- Filtering of peaks: Some methods filter out peaks that do not represent harmonic content or the leading voice. For this, they filter the peaks taking into account magnitude or sinusoidality criteria (Rao and Rao, 2010).
- Spectral magnitude normalization: The aim of this method is to minimize the influence of timbre on the analysis. Some examples of this technique consist in taking the log spectrum (Arora and Behera, 2013) or applying spectral whitening (Ryynänen and Klapuri, 2008).
- Computing instantaneous frequency: Some algorithms obtain the instantaneous frequency from the phase spectrum in order to refine the frequency and amplitude estimations of the spectral peaks (Dressler, 2011).

4. *Computation of the salience function*: At this step it is obtained the salience function, which is a multipitch representation containing the information about the salience of each pitch candidate at each time instant. Then, the peaks of this function are considered as possible candidates for the melody. There are several methods to compute the salience function. Some of them are presented below:

   - Harmonic summation: The salience of each pitch is calculated as the weighted sum of the amplitude of its harmonic frequencies. This is the approach employed by most of the algorithms (Salamon and Gómez, 2012).
   - Tone models: Some algorithms employ expectation maximization to fit a set of tone models to the observed spectrum (Goto, 2004). Then, the estimated maximum a posteriori probability (MAP) of the tone model whose fundamental frequency correspond to a certain pitch is considered as the salience of that pitch.
   - Other approaches: Other algorithms employ two-way mismatch (Rao and Rao, 2010), summary autocorrelation (Paiva et al., 2006) and pairwise analysis of spectral peaks (Dressler, 2005).

5. *Octave errors minimization*: One of the main problems of the salience function based methods is the appearance of “ghost” pitch values, whose fundamental frequency is an exact multiple (or sub-multiple) of the fundamental frequency of the pitch of the predominant melody. This issue may lead to what is called octave errors, in which an algorithm selects a pitch value located exactly one octave above or below the actual pitch of the melody. Below some approaches to reduce this kind of errors are explained:

   - Reducing the number of “ghost” pitch values in the salience function: In order to achieve this, Dressler (2011) examines pairs of spectral peaks potentially belonging to the same harmonic series and attenuates their summation in case there are many high amplitude spectral peaks with a frequency between the considered pair. Another approach proposed by Cancela (2008) attenuates the harmonic summation of a certain fundamental frequency f_0 if the mean amplitude of the components at frequencies $2kf_0, 3kf_0/2$ and $3kf_0$ is above the mean of the components at frequencies kf_0, attenuating this way the "ghost" pitch values whose f_0 is $1/2$, $2/3$ or $1/3$ of the actual f_0.

- Spectral smoothness: After smoothing the spectral envelope of the harmonics of the pitch candidates the salience function is recomputed. The candidates with an irregular envelope will be considered as "ghosts" that conduce to octave errors and will be attenuated (Klapuri, 2004).
- Pitch contours: This approach first groups the peaks of the salience function into pitch contours and then removes the "ghost" contours. This can be done by identifying duplicate contours, which will have the same shape but one octave apart, and filtering out the "ghost" taking into account contour salience and pitch continuity criteria (Salamon and Gómez, 2012).
- Continuity criteria: Finally, all methods reduce the octave errors in an indirect way by penalizing large jumps in pitch during the tracking step of the algorithm.

6. *Tracking*: Once the peaks of the salience function have been calculated, it is necessary to decide which peaks belong to the melody. This is a critic step and there are a lot of alternatives to perform it, almost one for each proposed algorithm. Most of them try to directly track the melody from the salience peaks, while others include a previous step in which peaks are grouped into continuous pitch contours. Some of the proposed tracking methods include: clustering (Marolt, 2005), heuristic-based tracking agents (Goto, 2004), Hidden-Markov Models (HMM) (Yeh et al., 2012), dynamic programming (Hsu and Jang, 2010) or filtering all the contours that do not belong to the melody instead of tracking it (Salamon and Gómez, 2012).

7. *Voicing Detection*: At this step it is decided when the predominant melody is present. It is usually applied at the end, although there are some exceptions. For instance, Salamon and Gómez (2012) use a threshold based on the salience distribution of pitch contours to remove non-salient contours before in order to filter out other non-melody contours. Other approaches include:

   - Using a per-frame salience-based threshold which can be fixed or dynamic (Paiva et al., 2006).
   - Incorporation of a silence model into the HMM tracking part (Ryynänen and Klapuri, 2008).

- Timbre based classification to determine the presence of the predominant instrument (Hsu and Jang, 2010).

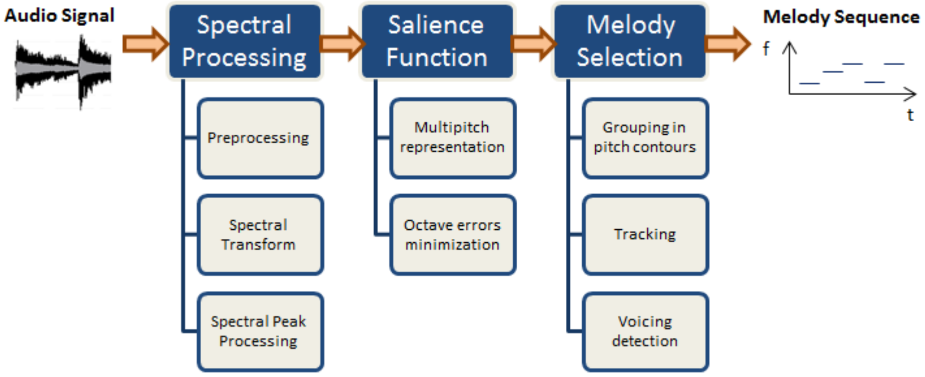


Figure 2.3: Block diagram of salience-based melody extraction algorithms.

Source-separation approaches: These algorithms first separate the predominant melody from the mixture and then perform an analysis in order to obtain the corresponding sequence of pitch values (Fig. 2.4). This approach has become popular in recent years due to the advances in audio source separation techniques. There are different approaches to enhance the melody signal:

- *Source/filter model*: This approach consists in modelling the contribution of the leading voice with a source/filter model. The source in voiced speech is the vibration of the vocal folds in response to airflow from the lungs and the filter is the vocal tract which is the tube from the glottis to the mouth. So, the voice can be represented as a signal produced by the vocal folds and filtered by a filter which modifies its spectrum. Using this model has sense because in most of the cases the predominant melody is the voice of a singer. Moreover, it can also be extended to some music instruments, for which the filter is then interpreted as shaping the timbre of the sound, while the source mainly consists in a generic harmonic signal driven by a fundamental frequency (Durrieu et al., 2010).

- *Harmonic/percussive separation*: This approach takes advantage of the variability of melody compared to more sustained chord notes using Harmonic-Percussive Sound Separation (HPSS). This algorithm was designed to separate harmonic from percussive elements present in a sound mixture by distinguishing between sources which are smooth in time (harmonic content) and sources smooth in frequency (percussive content). As a first step, it is used a changing window length for the analysis in order to separate the chords from the melody and percussive content. Then, once the accompaniment has been removed the algorithm is run again with its original configuration, in order to remove the percussion (Tachibana et al., 2010).

- *Repeating structure separation*: This technique is based on exploiting the fact that music accompaniment usually has a repetitive structure, while the predominant melody part has a higher grade of variability (Rafii and Pardo, 2013).

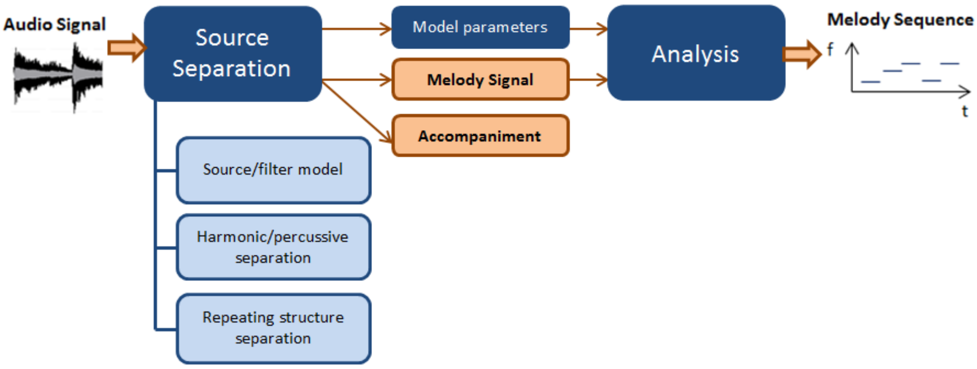


Figure 2.4: Block diagram of source-separation based melody extraction algorithms.

Alternative approaches: This group includes the algorithms that have a different approach from the previous. Some of the alternative proposals are presented below:

- *Data driven approach*: Poliner and Ellis (2005) propose to use machine learning in order to train a classifier to estimate the note directly from the power spectrum. They use a 256 feature vector to train a support vector machine classifier using training data labelled with 60 MIDI notes across 5 octaves.

- *Combination of monophonic pitch estimators*: Sutton et al. (2006) combine the output from two different monophonic pitch estimators using Hidden Markov Models.

This section has described the main characteristics of the different kind of existent melody extraction approaches. However, the provided references are only examples of algorithms that implement each specific feature. For further details the reader is referred to Salamon et al. (2014), in which it is exposed a comprehensive review of most of the techniques submitted to the Music Information Retrieval Evaluation eXchange (MIREX) from 2005 to date. MIREX is an annual campaign where different algorithms are evaluated against the same datasets in order to compare the quantitative accuracy of current state-of-the-art methods. So, it is a good starting point for reviewing because most of the algorithms with impact in the research community have been submitted to MIREX at least in one of its editions.

2.1.2.2 Relevant algorithms

This section presents a review of the algorithms that are relevant for the development of this master thesis which are mainly six. They represent the state-of-the-art and have been evaluated in the context of the MIREX initiative :

- **Dressler (2009):** First it obtains a multi resolution spectrogram representation from the audio signal by computing the STFT with different amounts of zero padding using a Hann window. So, it computes the STFT spectra in different time-frequency resolutions. For all spectral resolutions the length of the analysis window and the hop size are 2048 and 256 samples, respectively (assuming a 44.1 kHz sampling frequency). Then it is computed the magnitude and phase spectra. Next, the peaks of the spectra are selected using a simple magnitude threshold which is a fraction of the biggest peak magnitude of the current frame. The following step is to compute the instantaneous frequency (IF) for the selected peaks using the average of two different methods: phase vocoder and the method proposed by Charpentier (1986). After that, the spectral peaks are analysed in a pair-wise way with the aim of identifying partials with a successive harmonic number. Next, the identified harmonic peak pairs are evaluated following a perceptually motivated rating scheme and the resulting pitch strengths are added to a salience function with a frequency range between 55 and 1318 Hz. In this case, the salient pitches function is only the starting point for new tone objects. The actual estimation of tone height and tone magnitude is performed as an independent computation: harmonic peaks are added to existing tone objects and after a short time a timbre representation for that tone is established. The timbre will determine how much harmonic partials of the current frame will influence pitch and magnitude of the tone. This way the impact of noise and other sound sources can be decreased. At the same time the frame-wise estimated pitch candidates are processed to build acoustic streams. A rating is calculated for each tone depending on loudness, frequency dynamics, tone salience and tone to voice distance. Tones with a sufficient rating are assigned to the corresponding streams. Finally, the most salient auditory stream is identified as the melody.

- **Durrieu et al. (2010)**: They propose a signal model where the leading vocal part is explicitly represented by a specific source/filter model. This approach is investigated in the framework of two statistical models: a Gaussian Scaled Mixture Model (GSMM) and an extended Instantaneous Mixture Model (IMM). This source filter/model has a fixed number of possible fundamental frequencies and a fixed number of filters which represent the pronounced vowels or possible timbres. Moreover, it is characterized by the source spectra dictionary matrix and the filter spectra shape matrix. The source spectra dictionary is fixed and contains the spectrum for all the possible fundamental frequencies in a range from 100 to 800 Hz and it is computed using this Glottal Source Model. The model for the accompaniment is inspired by non-negative matrix factorization (NMF) with the Itakura-Saito divergence. The spectral model used both for the leading voice and the accompaniment is the STFT, but using a statistical approach. So, the Fourier Transform of each frame is approximated as a Complex Gaussian Variable with mean zero and with a diagonal covariance matrix whose coefficients are the Power Spectrum Density. Then, the parameters of the different models are estimated using the Maximum-Likelihood Principle, employing the Expectation-Maximization Algorithm for the GSMM and the Multiplicative Gradient Method for the IMM. After that, they define a Hidden Markov Model and use the Viterbi algorithm to smooth the melody line. Finally, in order to recognize silences they include a null spectrum state in the proposed Hidden Markov Model.

- **Fuentes et al. (2012)**: This algorithm is based on the Constant-Q Transform (CQT) of the mixture. It is a source separation based approach which relies on a Probabilistic Latent Component Analysis (PLCA) model. The accompaniment model is the classical PLCA equivalent to the Non-Negative Matrix Factorization and the melody model is a Shift-Invariant PLCA. The parameters of the models are estimated using the Expectation Maximization (EM) algorithm. Then, the melody is tracked using the Viterbi algorithm proposed by Durrieu et al. (2010). For the voicing detection step, first the temporal energy signal of the estimated melody is filtered with a 1/10Hz cut-off frequency low-pass filter, then a threshold manually set at -12dB is applied.

- **Hsu and Jang (2010)**: This algorithm is based on a trend estimation algorithm which detects the pitch ranges of a singing voice in each time frame. First, the singing voice is enhanced by considering temporal and spectral smoothness using the harmonic/percussive sound separation (HPSS) proposed by Tachibana et al. (2010). Then, the sinusoidal partials are extracted from the mixture by applying the MRFFT proposed by Dressler (2006). After that, the peaks are grouped, so that each peak corresponds to a partial. This stage consists of three steps: initial grouping, re-grouping, and refining. Then, they extract features from each partial in order to consider the natural differences between vocal partials and instrumental partials (e.g. vibrato and tremolo) and use a classifier to detect and prune instrumental partials . The next stage is to find a sequence of relatively tight pitch ranges where the F0s of the singing voice are present. First, harmonic partials are deleted based on the observation that the vocal F0 partial can only be the lowest-frequency partial within a frame. Then, they downsample the magnitudes of the partials by summing the largest peak values in the frames within a T-F block, which is a rectangular area whose vertical side represents a frequency range and horizontal side represents a time duration. Finally, they find an optimal path consisting of a sequence of T-F blocks that contain the largest downsampled magnitudes by using dynamic programming (DP).

- **Paiva et al. (2006)**: This is a multistage approach, inspired by principles from perceptual theory and musical practice. It comprises three main modules: pitch detection, determination of musical notes (with precise temporal boundaries, pitches, and intensity levels), and identification of melodic notes. The pitch detector is based on Slaney and Lyon (1993) auditory model, using 46.44 ms frames with a hop size of 5.8 ms. For each frame, a cochleagram and a correlogram are computed, after which a pitch-salience curve is obtained by adding across all autocorrelation channels. The pitch salience in each frame is approximately equal to the energy of the corresponding fundamental frequency. Unlike most other melody-extraction systems, they attempt to explicitly distinguish individual musical notes (in terms of their pitches, timings, and intensity levels). To do that, they first create pitch tracks by connecting pitch candidates with similar frequency values in consecutive frames. Each trajectory may contain more than one note, so it should be segmented in time. This is performed in two phases, namely frequency-based segmentation and salience-based segmentation. Finally, they identify the final set of notes representing the melody by means of a set of rule-based systems that attempt to extract the notes that convey the main melodic line among the whole set of detected notes. Moreover, they make use of some perceptual rules of sound organization such as harmonicity or common fate to eliminate "ghost" pitches.

- **Salamon and Gómez (2012):** This approach first applies a perceptually motivated equal loudness filter that enhances the frequencies to which the human listener is more sensitive. Then, they compute the STFT using a Hann window of size 46.4 ms, a hop size of 2.9 ms and a x4 zero padding factor. Next, the spectral peaks are selected by finding all the local maxima of the magnitude spectrum. The frequency and amplitude of the peaks is corrected by calculating the peak's instantaneous frequency. After that, they compute the salience function by harmonic summation using only the spectral peaks. The proposed salience function covers a range from 55 to 1760 Hz, quantized into 600 bins of 10 cents each one. Then, they create pitch contours from the peaks of the salience function using auditory streaming cues such as time-continuity or pitch continuity. These contours are sequences of pitch candidates which are characterized using the following features: pitch mean, pitch deviation, contour mean salience, contour total salience, contour salience deviation, length and vibrato presence. Next, for the voicing detection they define a threshold based on the mean salience of all contours in order to filter out those with a low salience. Moreover, they also include a procedure to delete "ghost contours" based on the fact that the correct contour will have a greater salience than its duplicate and will not represent a large jump, avoiding this way octave errors. Finally, they select the peaks belonging to the main melody from the remaining contours. Taking into account the creation and characterization of the contours, this is a simple task because in most of the cases there will be only one peak to choose. However, if there are still various contours, the melody is selected as the peak belonging to the contour with the highest total salience. In case there are not any contours present in the frame, it will be considered as unvoiced.

Algorithm	**Dressler**	**Durrieu**	**Fuentes**	**Hsu**	**Paiva**	**Salamon**
Approach	Salience based	Source separation	Source separation	Salience based	Salience based	Salience based
Preprocessing	-	-	-	Harmonic percussive separation	-	Equal loudness filter
Spectral Transform	MRFFT	STFT	CQT	MRFFT	Auditory model	STFT
Peak selection	Magnitude threshold + IF	Source filter model	PLCA	Vocal partial discrimination	Autocorrelation peaks	IF
Salience function	Pairwise comparison of spectral peaks			Normalized subharmonic summation	Summary correlogram	Harmonic summation
Octave errors minimization	Pairwise comparison			Pitch contours	-	Pitch contours
Tracking	Streaming rules	Viterbi smoothing	Viterbi smoothing	Global trend + dynamic programming	Multipitch trajectories + note deletion	Contour tracking + filtering
Voicing	Dynamic threshold	Energy Threshold	LPF + Energy Threshold	Classification	Salience valleys	Salience distribution

Table 2.1: Main features of a selected set of algorithms.

2.2 Evaluation

Attending to the importance that the user has in the process, MIR systems can be classified into two groups:

- *System-based MIR*: This includes all the research that includes only laboratory experiments carried out by means of a computer (e.g. evaluation of algorithms using quantitative metrics). These approaches typically have been focused on computational models whose aim is to describe the different aspects of how humans perceive the music, for instance by means of musical feature extractors or similarity measures. So, these approaches assume the existence of a "ground truth" against which the output of the different MIR algorithms can be evaluated.

- *User-centric MIR*: In contrast, this approach always involves the interaction of MIR systems with human subjects. As a consequence, the concept of "ground truth" becomes obsolete because of the subjectivity that the interaction with users supposes. That is, users will be conditioned by factors such as their music preferences, musical training or demographics.

This section reviews the evaluation task from both points of view. First, the current system-centered evaluation of melody extraction systems is presented. Then, the performance of the state-of-the-art algorithms is explained. Second, it is presented an introduction to user-centered evaluation and its motivation. Finally, some systems found in the literature that somehow involve users in the evaluation process are reviewed.

2.2.1 System-centered evaluation

Melody extraction algorithms are evaluated following the metrics proposed by MIREX, an annual campaign in which state-of-the-art algorithms are evaluated against the same datasets in order to compare them (Downie et al., 2005). Typically, melody extraction algorithms operate by frames, so the output usually consists of two columns. The first with the timestamps separated a fixed interval (e.g. 10 ms interval for MIREX) and the second with the fundamental frequency values representing the pitch estimation at each analysis frame. As explained before, melody extraction algorithms are expected to complete two stages: estimate when the melody is present (voicing detection) and estimate the correct pitch of the melody (pitch estimation). The evaluation of these two steps can be done independently by the incorporation of negative pitch values. That is, by putting a negative sign in the estimation algorithms can report a pitch value even for frames in which the algorithm estimated that the melody is absent (non-melody frames).

Given the output of an algorithm for an audio excerpt, the evaluation is performed by means of a comparison with a ground truth annotation, which has the same format that the algorithm's output, with the exception that non-melody frames are labelled with a zero pitch value. In order to be able to produce ground truth files of a specific songs it is necessary a multitrack recording of that song to run a monophonic pitch tracker on the solo melody track. After that, a graphical user interface such as Sonic Visualizer can be used to inspect the output and manually correct the errors that might appear.

Once the ground truth is available the evaluation is done comparing the output and the ground truth on a per-frame basis. For non-melody frames the algorithm is expected to indicate the absence of melody, and for melody frames the algorithm must return a pitch value within half a semitone (50 cents) of the value annotated in the ground truth.

2.2.1.1 Current metrics

This section presents the five quantitative metrics that are used in MIREX to evaluate the melody extraction algorithms accordingly to Salamon et al. (2014). The notation employed along this section is the following:

- $\boldsymbol{f}$ is the vector containing the pitch frequency sequence that represents the melody extracted by the algorithm (recall that the algorithms may report also negative values, so the absolute value of the output is used).
- $\boldsymbol{f^*}$ is the vector containing the ground truth corresponding to the previous sequence.
- $\boldsymbol{v}$ is the voicing indicator vector whose t element is $v_t = 1$ when a melody pitch is detected by the algorithm (that is, when the reported pitch value is major than zero).
- $\boldsymbol{v^*}$ is the corresponding ground truth voicing indicator.
- $\overline{\boldsymbol{v}}$ is the "unvoicing" indicator which can be defined for the frame t as:

$$\overline{v_t} = 1 - v_t$$

- $\overline{\boldsymbol{v}}^{\boldsymbol{*}}$ is the corresponding ground truth "unvoicing" indicator.

The different evaluation metrics are explained below:

- **Voicing Recall Rate:** It is the proportion of frames labelled as melody frames in the ground truth that are estimated as melody frames by the algorithm.

$$Rec_{vx} = \frac{\sum_t v_t v_t^*}{\sum_t v_t^*}$$

- **Voicing False Alarm Rate:** It is the proportion of frames labelled as non-melody in the ground truth that are estimated incorrectly as melody frames by the algorithm.

$$FA_{vx} = \frac{\sum_t v_t \overline{v}_t^*}{\sum_t \overline{v}_t^*}$$

- **Raw Pitch Accuracy:** It is the proportion of melody frames in the ground truth for which the estimated pitch is correct (i.e. within 50 cents of the ground truth).

$$Acc_{pitch} = \frac{\sum_t v_t^* T[M(f_t) - M(f_t^*)]}{\sum_t v_t^*}$$

where T is a threshold function defined by:

$$T[a] = \begin{cases} 1 & if\ |a| < 50 \\ 0 & if\ |b| \geq 50 \end{cases}$$

and M converts from a frequency in Hz to Cents with respect to an arbitrary reference frequency (which typically is A1 or 55Hz):

$$M(f) = 1200\, log_2 \left(\frac{f}{f_{ref}} \right)$$

- **Raw Chroma Accuracy:** It is as raw-pitch accuracy with the exception that both the estimated and ground truth pitch are mapped onto a single octave:

  $$Acc_{chroma} = \frac{\sum_t v_t^* T[\langle M(f_t) - M(f_t^*) \rangle_{12}]}{\sum_t v_t^*}$$

  where $\langle a \rangle_{12}$ denotes pitch values modulo 12.

- **Overall accuracy:** This measure combines the performance of the pitch estimation and voicing detection tasks to evaluate the overall system. It is defined as the proportion of all frames correctly estimated by the algorithm, where for non-melody frames this means the algorithm labelled them as non-melody and for melody frames the algorithm both labelled them as melody frames and estimates the correct pitch (i.e. within 50 cents of the ground truth):

  $$Acc_{ov} = \frac{1}{L} \sum_t v_t^* T[M(f_t) - M(f_t^*)] + \overline{v}_t^* \overline{v}_t$$

Furthermore, the performance of an algorithm typically is rated on an entire collection. This is carried out by averaging the per-excerpt scores for the different measures over all excerpts in the specific collection. Over the years, different research groups have annotated music collections for evaluating melody extraction in MIREX. In particular, the dataset employed during the development of this master thesis is composed of excerpts similar to the test collection showed in Table 2.2. Other collections used in MIREX are INDIAN08, which contains Indian classical vocal performances, and two different versions of MIREX09 with a signal-to-accompaniment ratio of -5dB and +5dB respectively.

Finally, some issues about the reliability of the evaluation of Audio Melody Extraction algorithms proposed by Salamon and Urbano (2012) are briefly reviewed. First, the annotation process of the ground truth is a very important step that must be done carefully. For instance, Salamon and Urbano (2012) identified a time offset between the algorithms output and the ground truth annotation in a music collection used for Audio Melody Extraction evaluation in MIREX. Taking into account that all the evaluation measures are based on a frame-by-frame comparison of the algorithm's output to the ground truth annotation, this caused

Description	**Description**	**Total play time (s)**
ADC2004	20 excerpts of roughly 20 s in the genres of pop, jazz and opera. Includes real recordings, synthesized singing and audio generated from MIDI files.	369
MIREX05	25 excerpts of 10-40 s duration the genres of rock, R&B, pop, jazz and solo classical piano. Includes real recordings and audio generated from MIDI files.	686
MIREX09	374 Karaoke recordings of Chinese songs.	10020

Table 2.2: Test collections.

a mismatch in all frames. Since melody pitch tends to be continuous, a very small offset may not be noticed. However, as the offset is increased, we observe a bad effect on the results.

Another issue is the use of clips instead of full songs. The collections used in the evaluations contain some very short excerpts, some only 10 seconds long. Moreover, these short clips primarily contain voiced frames, so the generalization of the results to full songs should be questioned. The studies carried out by Salamon and Urbano (2012) suggest that there is a correlation between the relative duration of the clip compared to the full song and the evaluation error. So, performance based on clips might not really predict performances of whole songs. However, how long a clip must be in order to reliably evaluate the system does not depend on how long is the song. In fact, the clip must be selected in such a way that it is representative of the whole song independently of its duration.

Concerning the current evaluation collections, Salamon and Urbano (2012) conclude that the collections ADC04, MIREX05 and INDIAN08 are not very reliable, because a larger proportion of the variability in the observed performances scores is due to the song difficulty differences rather than algorithm differences. So, they suggest to fuse these collections into a larger one to obtain more stable results. Finally, the MIREX09 collection is much larger than necessary. Moreover, all MIREX09 material consists of Chinese karaoke songs with non-professional singers. So, the evaluation results using this collection are not generalizable to other kind of music different from karaoke.

2.2.1.2 Performance of algorithms

To get a general idea of the performance of the algorithms it is only necessary to look at two evaluation measures: the raw pitch accuracy, which informs about how well the algorithm tracks the pitch, and the overall accuracy, which also reflects the quality of the voicing detection task. Moreover, the accuracy of all algorithms varies depending on the collection being analysed. The raw pitch accuracy over all collections for the state-of-the-art algorithms lies between 70-80% while the overall accuracy is in all the cases lower being between 65% and 75% for the best performing algorithms. The overall accuracy for the relevant algorithms for this master thesis is shown in Fig. 2.5. All these algorithms were submitted to MIREX with the exception of the one proposed by Fuentes et al. (2012). Moreover, for the algorithm proposed by Paiva et al. (2006) there are only results for the MIREX05 collection. Finally, just mention that the average value showed in this case is not weighted by the number of files of each collection.

Salamon et al. (2014) presented a comprehensive comparison of the different algorithms submitted to MIREX since 2005. They conclude that the performance has not improved much over the last 4 years. This highlights an important limitation of the MIREX evaluation campaign: since the collections employed for the evaluation are kept secret, it is very hard for researchers to learn from the results.

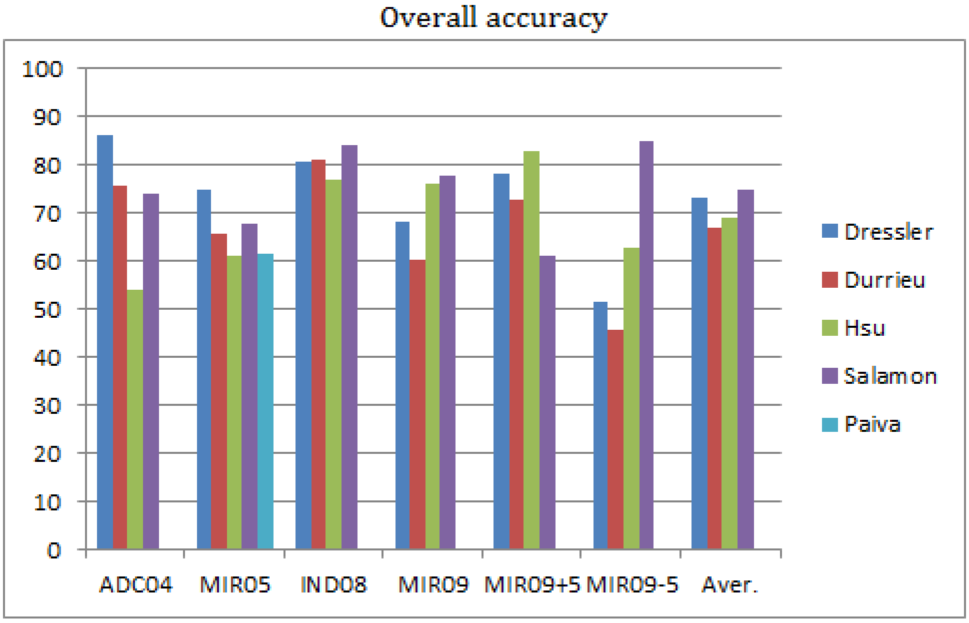


Figure 2.5: Overall accuracy for different collections of the relevant algorithms for this master thesis.

2.2.2 User-centered evaluation

This section emphasizes the importance of the user in the evaluation of MIR and reviews some MIR algorithm that involve the user in the evaluation process.

2.2.2.1 Introduction and motivation

The evaluation of a retrieval system is a crucial aspect to develop and improve the systems. Up to now, the evaluation approach followed by current MIR systems has been mainly system-based (e.g. melody extraction task). This has been motivated because MIREX, which is the main evaluation campaign in the field, only employs system-centered measures, such as *accuracy* for the classification tasks, *average precision* for retrieval tasks and variations of *precision/recall* for detection tasks.

These system-centered measures are employed to give a score to the systems that determines how well they perform. It is commonly assumed that systems with better scores are actually perceived as more helpful by the users and consequently are expected to provide more satisfaction. Furthermore, researchers usually compare systems using the difference between these measures (Urbano et al., 2012).

That is the case of the melody extraction task in which we usually use the overall accuracy to determine if a system A is better or worse than a system B, assuming that the one with higher accuracy is better. However, it can be demonstrated that this is not always true. It can be proved that systems with a higher overall accuracy could bring less satisfaction to the users than others with a lower overall accuracy due to the nature of the different errors involved in the task (Chapter 3).

Since the ultimate goal of evaluating a MIR system is to characterize the usage experience of the users who will employ it, it is necessary to study the user satisfaction distributions and research in new metrics closer to the opinion of the users. Unfortunately, there are several problems when characterizing the experience of the users because including real users in experiments is expensive and complex. Furthermore, there are some ethical issues to consider such as the privacy and it is difficult to reproduce experiments that involve human subjects, so it is hard to compare systems across research groups (Urbano et al., 2013).

However, as the fundamental aim of MIR systems is to help users in searching music information, the evaluation of MIR systems should begin to move towards a user-centric approach. Although until now the user has been completely disregarded, there is a growing awareness of the importance of taking into account the opinion of the users (Hu and Liu, 2010) (Schedl et al., 2013).

2.2.2.2 Similar work

According to Schedl et al. (2013) the users may be relevant for the evaluation of the systems in different grades. Below some MIR algorithms that involve user-centered evaluation are presented, grouped by the extent of involvement that the users have in the task:

- **Precompiled user-generated datasets**: These techniques are evaluated on data generated by the users, but do not request feedback from them during the evaluation experiments.
  - Knees et al. (2007) proposed an automatic search-engine for large-scale music collections that can be queried by natural language. Since their goal was to develop a natural language search engine for music,

the evaluation was carried out by using “real-world” queries (phrases used by people to describe music and which are likely to be used when searching for music). In particular, they utilized the track specific tag information provided by Last.fm.

  - Lee and Lee (2007) developed a music recommendation system called C2_Music, which utilizes the users’ demographics and behavioural patterns and the users context. The evaluation of its performance relies on datasets of listening histories.
  - Xue et al. (2009) proposed a collaborative personalized search which considers not only a similarity factor among users for defining group user profiles and global user profiles, but also the specialities of each individual. The search engine identifies the users by a unique id (GUID). For evaluation, they randomly selected a corpus of web search data from 1000 users who have made at least 100 queries each one.

- **User response to a single question**: In this cases, the participation of the users in the evaluation consists in answering a single specific question. Although in this case the users are asked a question, their individual properties such as expertise or familiarity with the music items involved in the experiment are usually neglected. This approach could be correct for some cases, but such highly specific evaluation settings are not able to provide answers to other questions. For instance, more of the proposed cases exposed below fail at offering explanations for the suitability of the musical pieces under study.

  - Liu et al. (2009) proposed a music recommendation system based on the heartbeat of the user. In this case, during the evaluation the users are only asked if the tempo of the selected music is synchronized with their heartbeat.
  - The aim of the systems proposed by Biehl et al. (2006) and Moens et al. (2010) is to recommend music based on the pace of the user when doing sport. Moens et al. (2010) carried out an experiment with a total of 33 participants. After the experiment, the users filled in a survey asking their personal jogging preferences and their experience with the system, that is, if the tempo of the selected music was synchronized with their pace.
  - Kaminskas and Ricci (2011) proposed a context aware recommendation system which selects music content that fits a place of interest (POI). To address this problem they used emotional tags attached by a users’ population to both music and POIs. The evaluation was carried out by 10 users in total performing 154 evaluation steps, that is, each user considered on average 15.6 POIs and the music suggested for these POIs. During the evaluation the users were only asked the question of whether the music suggested is suited for the particular POI or not.

- **Multifaceted questionnaire**: These techniques try to solve the limitations of the "single-question" approach by asking more questions related to the users or their context:

  - Pauws and Eggen (2002) proposed the system Personalized Automatic Track Selection (PATS) to create playlists which suit a particular *context-of-use*, that is, the real-world environment in which the music is heard. It uses a dynamic clustering method in which songs are grouped based on their attribute similarity. During the evaluation they conducted an experiment with 22 participants, asking them to rate the resultant playlist. Moreover, it was conducted a post-experimental interview in order to gather supplementary findings on the perceived usefulness.
  - Pauws and van de Wijdeven (2005) propose "SatisFly", an interactive playlist generation system in which the user can tell what kind of songs should be contained in what order in the playlist, while navigating through a music collection. They evaluate the system by conducting an experiment with 24 participants. Each participant rated the generated playlist by measures such as the playlist quality, time spent on the task, number of button presses in accomplishing the task, perceived usefulness and ease-of-use.
  - Vignoli and Pauws (2005) conduct an user evaluation experiment with 22 participants to assess the "similar song" function of the music recommendation system E-Mu jukebox. In addition to the measures collected by Pauws and van de Wijdeven (2005) they also measured the order of participant's preference among three different systems.
  - Firan et al. (2007) proposed recommendation algorithms based on tag user profiles, and explore how collaborative filtering recommendations based on these tag profiles are different from those based on song profiles. They evaluate the proposed algorithms with 18 subjects who were asked to install a desktop application to extract their user profiles accordingly to the music present in their desktop. For each of the algorithms they collected the top-10 recommended items. Then, for each of the recommended tracks, the users had to provide two different scores: one measuring how well the recommended track matches their music preferences ([0] - *I don't like this track*, [1] - *I don't mind listening to this track*, [2] - *I like the track*) and one reflecting the novelty of the track ([0] - *I already know this track*, [1] - *I know something about this track, e.g. I know the artist or I heard the track on the radio, but I do not remember the name*, and [2] - *This track is really new for me*).
  - Bogdanov and Herrera (2011) consider distance-based approaches to music recommendation, relying on an explicit set of music tracks

provided by the user as evidence of his/her music preferences. During the evaluation step, a total of 19 voluntary subjects were asked to provide their respective preference sets and additional information, including personal data (gender, age, interest for music, musical background), and also a description of the strategy and criteria followed to select the music pieces. Then, a questionnaire was given for the subjects to express different subjective impressions related to the recommended music: A "familiarity" rating ([0]-*Absolute unfamiliarity*, [1]- *Feeling familiar*, [2]-*Knowing the artist* [3]- *Knowing the title* and [4]- *Identification of artist and title*). A "liking" rating which range from [0]-*Dislike* to [4]-*Like*. A rating of "listening intentions" measured preference with a similar range. Finally, a "give-me-more" rating allowing just [1] or [0] to respectively indicate a request for, or a reject of, more music like the one presented. The users were also asked to provide title and artist for those tracks rated high in the familiarity scale.

- Urbano (2013) conducted an experiment to map system effectiveness of melody similarity onto user satisfaction. Subjects were presented with different examples, each containing a query clip and two ranked lists of five results each, as if retrieved by two different audio melody similarity systems A and B. Then they had to listen to the clips and select one of the following options: system A provided better results, system B did, they both provided good results or they both returned bad results. Moreover, the questionnaire also includes a question to propose comments or suggestions about the experiment. User answers about the 22.074 relevance judgments of the ground truth across the 439 queries from the last editions of MIREX were collected via crowdsourcing.

Interestingly, all these studies were carried out on the context of the task of music recommendation or playlist generation, with similar design and scale, with the exception of the experiment conducted by Urbano (2013). This is an evidence of the fact that user-centered evaluation has not been well adopted and has limited influence in the MIR community.

Chapter 3

USER SURVEY

This chapter presents the procedure followed to design the experiment. First, it is described the preliminary perceptual analysis of the different kind of errors. Then, some hypotheses related to the perception of the melodies are proposed. Finally, it is presented the survey design and the obtained results.

3.1 Perceptual Analysis of Errors

The aim of this survey is to characterize real errors that current algorithms made, not any mistake that might come to our mind. So, in order to do a realistic experiment we have selected a dataset of 30 representative songs and the corresponding outputs from 6 different algorithms, which were kindly provided by their authors Dressler (2009), Durrieu et al. (2010), Fuentes et al. (2012), Hsu and Jang (2010), Paiva et al. (2006) and Salamon and Gómez (2012).

First, we have to synthesize the outputs in order to be able to listen to them. For the synthesis we have used a mixing of sinusoids containing five harmonics of the fundamental frequency value at each specific time instant[1]. The quality of the sound achieved is not the best, but it is enough for our purpose. Moreover, by this way there is no quantization in frequency. For example, if we had used MIDI we would have been subjected to a specific set of notes spaced one semitone, which is not appropriate for this experiment due to the fact that the frequency is a continuous variable. Moreover, current evaluation metrics use a tolerance of half a semitone, so in order to do comparisons with them, we should have at least the same accuracy.

Then, we listened carefully to all the outputs in order to formulate hypotheses that represent my ideas about how the different errors affect the perceived quality of the extracted melodies, analysing at the same time the spectral content of the original sounds to look for possible explanations of the mistakes. The terminology used to name the main kinds of errors is the following:

- Voicing errors: Errors detecting when the melody is present or not.

[1] For this purpose, we have used Matlab

  - False Alarm: Labelling a frame as containing the predominant melody when it is not present in the ground truth.
  - Missing Frame: Labelling a frame as non-containing the predominant melody when it is present in the ground truth.

- Pitch errors: Errors in the pitch estimation.
  - Octave errors: Selecting a pitch value which is exactly a integer number of octaves above or below the correct pitch of the melody. That is, the chroma of the pitch is correct.
  - Errors with other pitches: Pitch errors that also estimate bad the chroma.

3.2 Hypotheses

After listening to the outputs of the different algorithms for the proposed dataset, we formulated some hypotheses taken into account our subjective opinion about the quality of the extracted melodies, so these assumptions rely only on our own perception. In this section, we will explain the most significant ones.

3.2.1 Voicing Errors

First, the impact of a false alarm will depend on its time position with respect to the melodic context. For example, if there is a long silence in the melody, a false alarm in the middle will have a bigger impact than if it happens in a section of the melody where there is a high note density or at the end of a long note. Moreover, the pitch detected by the false alarm will have an important role, because adding a new note or a noise will sound worse than for instance lengthening a note. These considerations can be summed up in the following hypothesis:

Hypothesis 1 *The impact on the perceived quality of a false alarm depends both on its time position and on its pitch relative to the melodic context.*

Similar conclusions can be reached about the missing frames. For instance, shortening a note by missing some frames at the beginning or at the end of it may be overlooked by the users more easily than missing a full note or splitting it in two parts by a silence.

Hypothesis 2 *The impact on the perceived quality of a missing frame depends both on its time position and on the missed pitch relative to the melodic context.*

3.2.2 Pitch Errors

First, the continuity in pitch is an important factor for the perceptual evaluation of pitch. The more clear example is the effect of adding a kind of glissando when moving from one note to the following one. In this case, the quality of the extracted melody is better than if the wrong pitch is random, because despite the glissando the essence of the original melody is preserved. The same will occur in case there is a glissando in the original melody but not in the extracted one. Another example of the importance of the continuity in pitch and its relation to the melodic context is the fact that lengthening a note at the cost of shortening the following or vice versa will be less noticeable than a random pitch error.

Hypothesis 3 *The impact on the perceived quality of a pitch error will depend both on its time position and on its pitch relative to the melodic context.*

Moreover, the perceived quality may be inversely proportional to the frequency distance between the original pitch and the erroneous one. That is, if there is a short distance the quality will be better than if the pitch is far apart from the correct one.

Hypothesis 4 *The impact on the perceived quality of a pitch error will be proportional to the frequency distance between the correct fundamental frequency and the extracted one.*

Furthermore, the octave errors may be less annoying for certain users than the errors with other pitches because the octave errors preserve the chroma dimension of the pitch sensation.

Hypothesis 5 *The impact on the perceived quality of an octave error will be less than that of a generic pitch error.*

Moreover, there are algorithms that sometimes extract the melody played by other instrument different from the predominant one, being this error a false alarm or a pitch error. However, in some cases it is not clear which is the predominant melody, so confusing the instruments may not be considered an error by some users. In any case, if the algorithm extracts the melody of a secondary instrument, this will be less annoying than extracted noise or something that has nothing to do with the original excerpt of polyphonic music.

Hypothesis 6 *Extracting the melody from another instrument different from the one playing the predominant melody has less impact in the perceived quality than other noisy false alarms or pitch errors.*

Finally, there are other considerations that should be taken into account when evaluating the pitch errors. For instance, the current quantitative metrics use a tolerance of 50 cents, but there are some pitch estimations that in spite of fulfilling this requirement have not a good perceived quality. For instance, if the predominant melody has a deep vibrato, a bad reconstruction of it may be quite annoying even if the pitch tolerance is respected.

Hypothesis 7 *If the predominant melody has a deep vibrato, the perceived quality of the extracted melody will be conditioned by the correct reconstruction of the vibrato.*

3.2.3 Time continuity

Another important issue that affects all kind of errors is the continuity in time. Until now, the quantitative evaluation is based on a per frame comparison between the extracted pitch and the ground truth. However, this approach may be inappropriate for an user-centered evaluation. For instance, suppose the two cases shown at Fig. 3.1. Both (a) and (b) have the same number of erroneous frames, however when listening to (a) the sensation is that the first half of the excerpt is good and the last half is bad, while when listening to (b) the melody may be totally unrecognisable due to the distribution of the errors. So, if we call each group of consecutive incorrect frames a grouped error:

Hypothesis 8 *The perceived quality of the extracted melody will depend both on the distribution of the incorrect frames and on the number of grouped errors.*

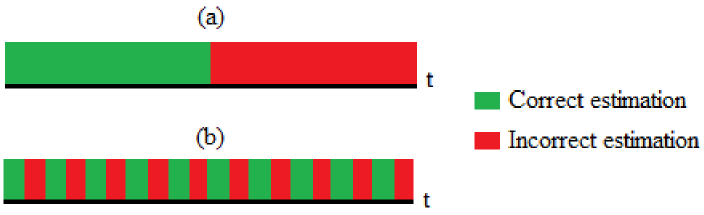


Figure 3.1: Importance of time continuiy and groupping of erroneous frames.

3.3 Survey design

After formulating the hypotheses, the next step is to test them taking into account the opinion of the users by means of a survey. The designed form has two parts: the first one collects some personal information and the second one includes the perceptual evaluation of the extracted melody from 20 excerpts. The full questionnaire can be found in Appendix A.

Regarding the personal information, although the survey is anonymous people are encouraged to provide an alias and an email in order to contact them if necessary. They are also asked their age, just to ensure their hearing health is well. Moreover, they have to answer some questions about their musical background including their training, the instruments they play and their singing skills, among others.

In the second part, first it is explained the concept of predominant melody. Then, people have to grade on a scale of 0 to 5 the extracted melody for 20 excerpts, where:

- 0 means the quality of the extracted melody is very bad compared to the ground truth.

- 5 means the extracted melody is as good as the proposed ground truth.

For each excerpt, the users have to listen to three different audio clips:

- *Original* is the excerpt representing the original recording of polyphonic music.

- *Ground Truth* is the synthesized sequence of the frequency values that best represent the predominant melody of the original excerpt.

- *Melody* is the synthesized sequence of the frequency values estimated by different melody extraction algorithms.

This decision is motivated by the hypothesis 6. This way, the users can compare the content of the original audio with the extracted melody. Otherwise, it would have been enough listening to the ground truth and the extracted melody. Moreover, the different excerpts have been presented randomly to the users, so the order does not affect the result. Finally, apart from listening and grading the clips, they can also comment their criteria or thoughts about each excerpt, although this is not compulsory. This way, we can gather information about the justification of the decisions made by the users. Moreover, they can also offer global opinions about the full experiment.

The main limitation when designing this survey is the fact that we can only use errors that appear in our dataset, so the variety of cases is limited. Moreover, the survey should be short because otherwise the subjects would have not properly fulfilled it for free. This is the reason for including only 20 short excerpts. These excerpts have been selected with the aim of testing the previous hypotheses, so each one contains only a significant grouped error which is useful to prove one of the hypotheses. In order to isolated the different errors, the original mistake made by a specific algorithm is mixed with the values of the ground truth, in order to provide a melodic context for the test. This has been done adding 2 seconds of ground truth frames at the beginning and at the end of the excerpt, being the error located in the middle, with the exception of the fragments related to the hypothesis 7 in which it has been added only one second for designing reasons. Below we present the characteristics of the selected audio clips.

3.3.1 Voicing Errors

In order to test the hypothesis 1 related to the impact of the false alarms, we are going to focus on the fact that lengthening a note has less impact on the perceived quality than adding a random noise. This way, we can demonstrate that the time position of the error relative to the melodic context influences the perception, because in this case the errors must be located at the beginning or end of a note to go unnoticed by the users. Moreover, we can also prove that the pitch of the false alarm relative to the melodic context is relevant to determine its impact on the perceived quality, because to give the sensation that a note is lengthened the pitch of the false alarm must be equal to the one of that note.

To measure the perceived quality when a note is lengthened we have chosen four cases with increasing durations which are representative of the events that occur in our dataset (Fig. 3.2). This way, we can map the perception of the user to the duration of the error. Moreover, we can find out at what point the user realizes that the note has been lengthened. In all the cases the duration of the note is increased by some false alarms at the end of it changing its offset. Although we might assume that if we alter the onset the results will be similar, to affirm this we should carry out another experiment. Finally, we have included a case of a random false alarm with an intermediate duration in order to know how the users penalize this error in comparison to the previous ones (Fig. 3.3).

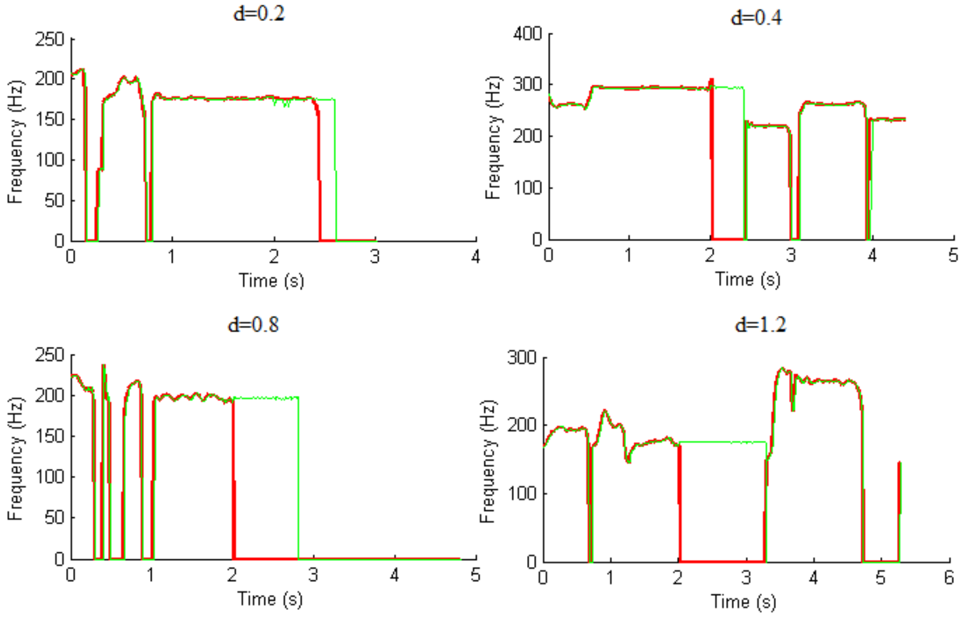


Figure 3.2: Pitch contours of excerpts with false alarms that lengthen a note. The duration (d) of each error is indicated in seconds. The red line represents the ground truth and the green line the extracted melody.

A similar approach will be followed to test the hypothesis 2 related to the missing frames. We are going to focus only on the fact that shortening a note may be overlooked by the user more easily than missing a full note. The chosen excerpts and their duration are shown in Fig. 3.4. As we can see, in this case we have include three cases in which the notes are shortened by putting forward the offset. However, this time we have also included a excerpt in which the onset of the note is delayed in order to check of this affects the perceived quality (excerpt with d=0.5s). Finally, we have included a clip in which a full note is missing (Fig. 3.5).

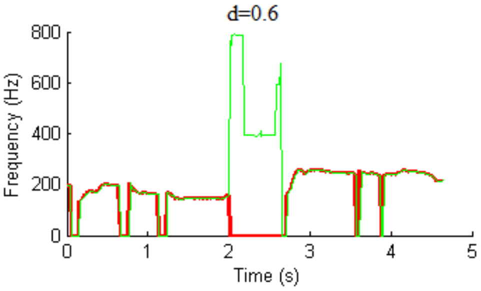


Figure 3.3: Pitch contour of a excerpt with noisy false alarms. The duration of the error is 0.6 s. The red line represents the ground truth and the green line the extracted melody.

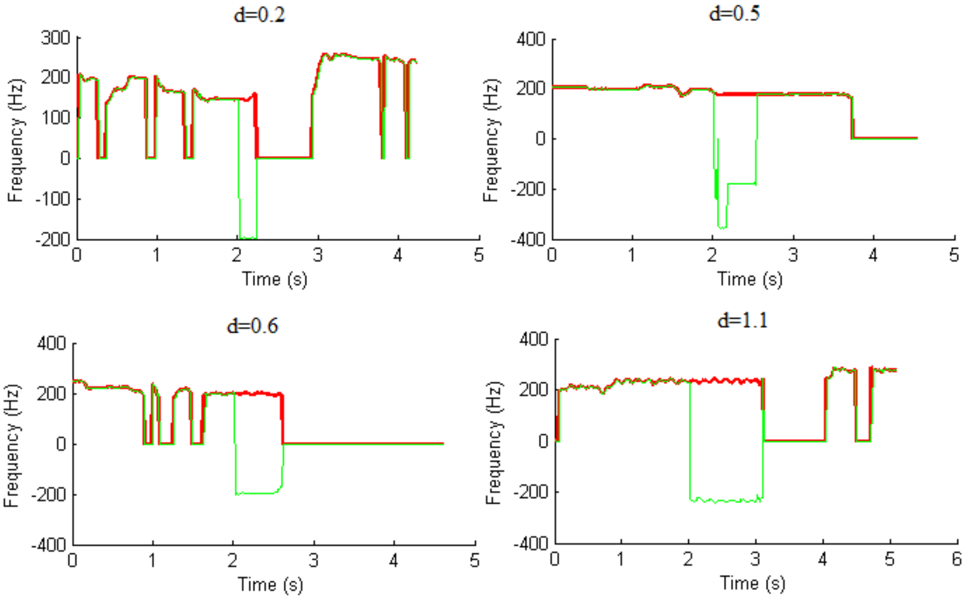


Figure 3.4: Pitch contours of excerpts with missing frames that shorten a note. The duration of each error is indicated in seconds. The red line represents the ground truth and the green line the extracted melody.

3.3.2 Pitch Errors

First, in order to test the hypothesis 3, which is about the relationship between the pitch errors and the melodic context, we have selected an excerpt in which the extracted melody contains a glissando that is not present in the original excerpt (Fig. 3.6). This decision is motivated by the fact that the presence of the glissando involves both a pitch and a time position relation between the incorrect frames and the melodic context. Furthermore, we have selected two excerpts with generic pitch errors without any specific relation to the melodic context. These errors have exactly the same duration than the glissando error (d=0.3s). Moreover, the average

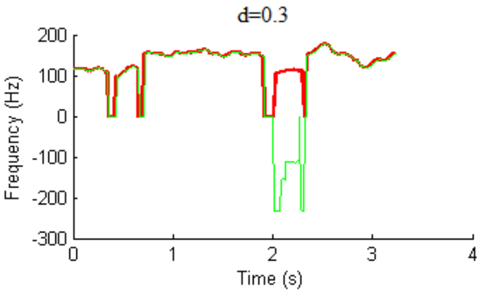


Figure 3.5: Pitch contour of an excerpt with a full note missing. The duration of the missing note is 0.3 s. The red line represents the ground truth and the green line the extracted melody.

distance between the ground truth pitch and the estimated one is a bit more than one octave for the first excerpt(~ 14 semitones) and a bit less than half an octave for the second excerpt (~ 5 semitones). This way, we can also test the hypothesis 4, which is related to the influence that the frequency distance between the pitch errors and the ground truth has on the perceived quality. So, we expect that the glissando error will have a higher grade than any of the generic errors to prove hypothesis 3. Moreover, the error shown at Fig. 3.7(a) should have a grade higher than the one at Fig. 3.7(b) to prove the hypothesis 4.

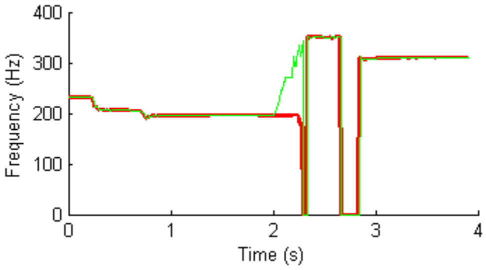


Figure 3.6: Pitch contour of an excerpt containing errors with other pitches including a glissando. The red line represents the ground truth and the green line the extracted melody.

Then, in order to prove the hypothesis 5 related to the octave errors we have selected two fragments, one containing an octave error and another one containing an error with other pitch (Fig. 3.8). Both have a duration of d=0.7s, which should be enough to perceive easily the octave error. To prove the hypothesis the grade of the octave error should be higher than the one of the error with a generic pitch.

Next, in order to prove the hypothesis 6 related to the extraction of the melody played by accompaniment instruments we have choose an excerpt in which it is extracted the melody played by a saxophone instead of the predominant singing voice. In this case there are both pitch errors and false alarms in order to take a duration enough to capture the 'essence' of the saxophone.

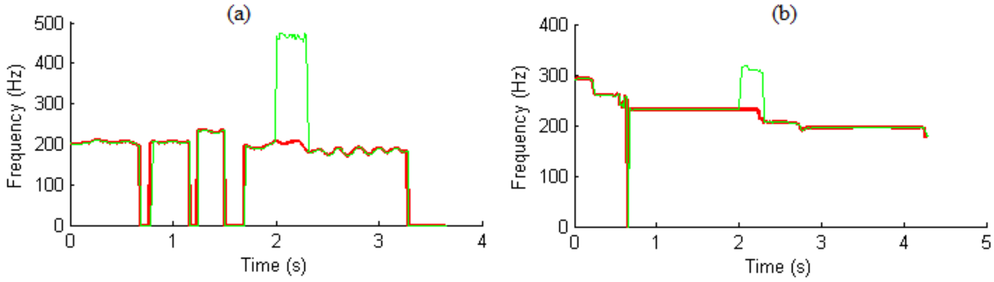


Figure 3.7: Pitch contours of excerpts containing errors with other pitches. In the excerpt (a) the erroneous pitch is “far” ($\sim$ 14 semitones) from the frequency of the ground truth and in the excerpt (b) it is “near” ($\sim$ 5 semitones). The red line represents the ground truth and the green line the extracted melody.

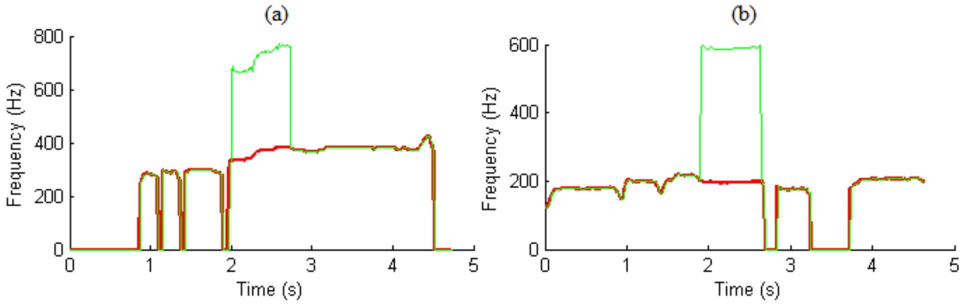


Figure 3.8: Pitch contour of an excerpt containing an octave error (a) and an excerpt containing an error with other pitch (b). The red line represents the ground truth and the green line the extracted melody.

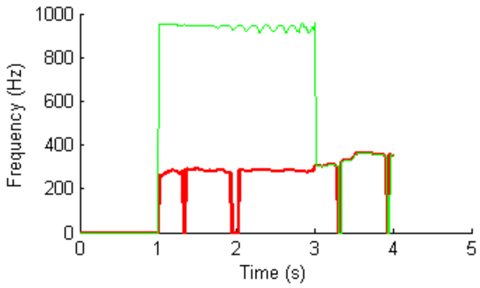


Figure 3.9: Pitch contours of an excerpt containing a pitch error in which the algorithm selects the melody played by an instrument of the accompaniment instead of the predominant one. The red line represents the ground truth and the green line the extracted melody.

Finally, in order to prove the hypothesis 7 related to the reconstruction of the vibrato, we have selected two excerpts, one with a good reconstruction and another one with a bad reconstruction (Fig. 3.10). Both excerpts have a few incorrect frames that will go unnoticed by the users, because the aspect that captures the user's attention here is the vibrato. Moreover, in order to facilitate the user to focus his/her attention on the vibrato, it has been added only a second of ground truth frames at the beginning and at the end which is enough since the melodic context is not particularly important to this hypothesis.

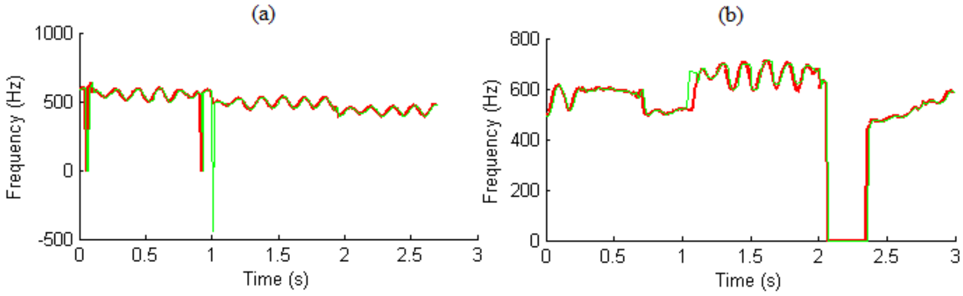


Figure 3.10: Pitch contours of two excerpts with a good (a) and a bad (b) reconstruction of the vibrato.

3.3.3 Time Continuity

To test the hypothesis 8 about the time continuity we are going to focus on the pitch errors. We have chosen two excerpts with the same number of incorrect frames but with a different distribution. In the first case, there is only one grouped error while in the second case the frames are grouped into two errors (Fig. 3.11). To prove the hypothesis the grade given to the excerpt (a) should be higher than the one given to (b). We may suppose that if we had selected other type of errors such as false alarms the results would have been the same, however in order to be rigorous we should carry out another experiment to extract that kind of conclusions. However, as we explained before the amount of excerpts that we can include in the survey is limited.

3.4 Survey results

Finally, 26 subjects aged between 56 and 23 years participated in the survey. Moreover, 20 subjects have a musical background, 5 subjects have no musical background and 1 subject did not answer the personal questions. We considered that a subject has musical background if he/she has answered yes to any of the questions proposed in the personal questionnaire (musical training, playing an instrument or sing). If we analyse the answers by groups, there are not big differences in the results. Furthermore, considering only the answers provided

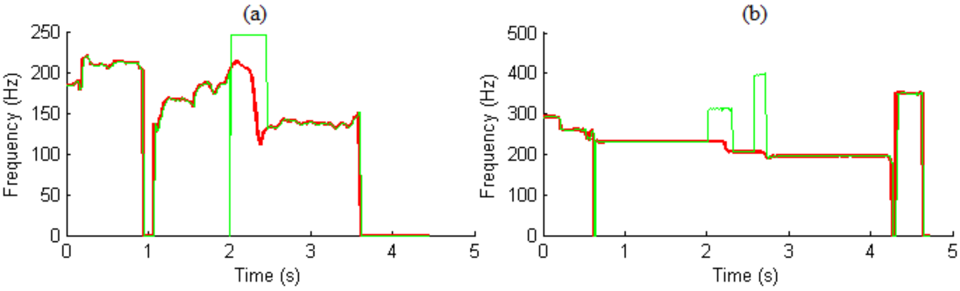


Figure 3.11: Pitch contours of two excerpts containing one (a) and two (b) grouped errors with other pitches. The red line represents the ground truth and the green line the extracted melody.

by 5 subjects may not be enough representative of the population without musical background. So, the results explained in this section have been calculated taken into account all the answers without any distinction.

For all the excerpts the analysis includes the median, the average, the standard deviation and the mode. Moreover, we have also included box plots with the results. On each box, the central mark is the median (red), the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually (+). In addition, in Appendix B we have also included the percentage of answers for each grade in order to analyse their distribution along the scale (0 to 5). In most of the cases, the distribution of answers is asymmetric or skewed, and we are using ordinal data. So, the most important measure in this case is the median and it is the one that we will take into account first when making decisions.

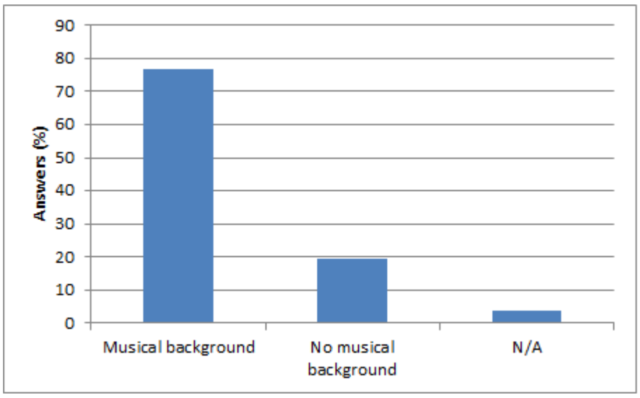


Figure 3.12: Percentage of answers grouping them accordingly to the musical background of the participants.

3.4.1 Voicing Errors

First, the results of the excerpts containing false alarms are shown in Table 3.1 and in Fig. 3.13 We can appreciate how the lengthening of the note goes unnoticed by the users until the error of duration d=0.8s for which the median is 4,5. Moreover, looking at the mode we can affirm that also this error goes unnoticed for most people. Actually, if we disregard the 0 grade of a subject who did not agree with the proposed ground truth, the median will be 5. Furthermore, if we compare the result of the excerpt of duration d=1.2s with the one of the random false alarm we can check that it has half the duration and it has also half the grade, which is totally opposite to the evaluation that we could achieve with current metrics (Table 3.2). Regarding the feedback, 9 subjects did a commentary about the duration of the notes, but although they perceive this issue they still grade them with a 4 or even a 5. So, the hypothesis 1 is true.

Type	Random	Lengthening a note			
Duration (s)	0.6	0.2	0.4	0.8	1.2
Median	2	5	5	4,5	4
Average	2,2	4,7	4,7	4,3	4,2
STD	0,9	0,5	0,5	1,0	1,0
Mode	2	5	5	5	5

Table 3.1: False alarms results.

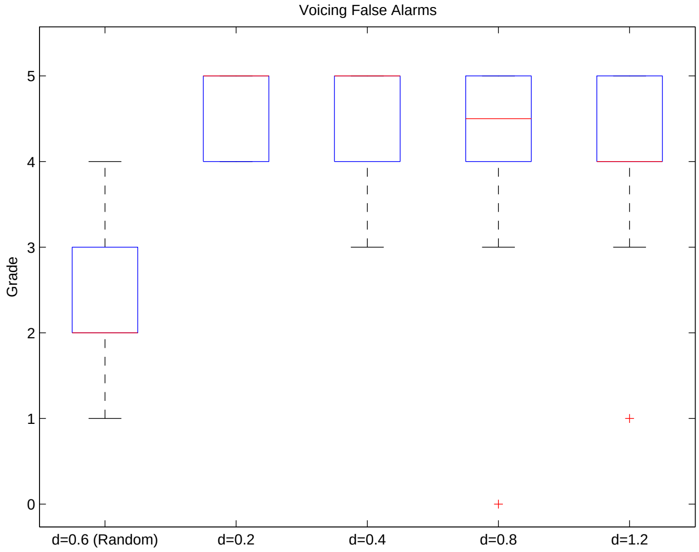


Figure 3.13: User ratings for the excerpts containing false alarms.

Type	Random	Lengthening a note
Duration (s)	0.6	1.2
Median	2	4
OA	86	76

Table 3.2: Overall Accuracy vs. Perceptual grade for excerpts containing only false alarms.

The results for the excerpt containing only missing frames are shown in Table 3.3 and in Fig. 3.13. We can see how shortening a note may go unnoticed by the users, but the duration of the error must be shorter than in the case of the false alarms lengthening a note. Furthermore, the difference between missing a full note and shortening a note is not as big as in the previous case with the random noise, but it is still significant (Table 3.4). So, the hypothesis 2 is true. Regarding the feedback, 11 subjects commented ideas about the tempo issues for the excerpts shortening a note and 7 commented the missing full note. Finally, it is important to notice that shortening the onset (excerpt with d=0.5s) has a higher penalization than shortening the offset. So, it is worth to carry out more experiments taking into account the differences between perceiving onsets and offsets.

Type	Full note	Shortening a note			
		Offset			Onset
Duration	0.3	0.2	0.6	1.1	0.5
Median	4	5	4,5	4	4
Average	3,8	4,7	4,3	3,7	4
STD	0,7	0,5	0,9	0,8	0,6
Mode	4	5	5	4	4

Table 3.3: Missing frames results.

Type	Full note	Shortening a note
Duration (s)	0.3	0.6
Median	4	4.5
OA	91	79

Table 3.4: Overall Accuracy vs. Perceptual grade for excerpts containing only missing frames.

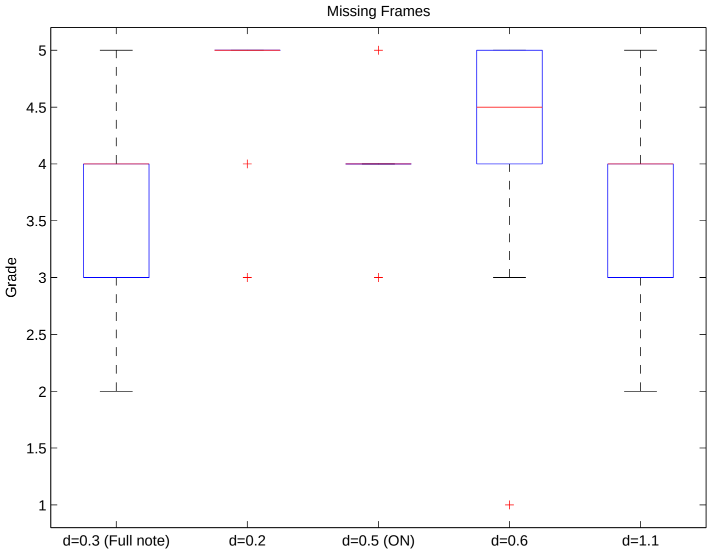


Figure 3.14: User ratings for the excerpts containing missing frames.

3.4.2 Pitch Errors

First, the results for the different excerpts containing only errors with other pitches are shown in Table 3.5 and in Fig. 3.15. We can observe that the median of the excerpt containing the glissando is one point higher than the median of the other two excerpts containing generic pitch errors denoted as "Far"(14 semitones apart from the ground truth) and "Near"(5 semitones apart from the ground truth). However, the excerpts with different frequency distance from the ground truth have exactly the same median. These three excerpts have exactly the same Overall Accuracy (93%), so the hypothesis 3 is true, but the hypothesis 4 can not be confirmed, because apparently the frequency distance has not enough influence in the perceived quality. Although looking at the edges of the box it seems that the "Near" error has received higher grades, it would be premature to drawn conclusions without more experiments because the difference between the grades is not as high as expected. Moreover, 6 subjects commented that there was an octave error which means that they are not able to distinguish properly between octave errors and generic pitch errors. Finally, 6 subjects commented the addition of the glissando.

Moreover, the inclusion in the experiment of the excerpt containing the melody of the saxophone instead of the voice (Other instrument) was not appropriate for these survey, because most people associate the ground truth with the perfection and directly grade this excerpt with a zero. So, the proposed task is not the most appropriated to demonstrate the hypothesis 6. Nevertheless, 6 subjects recognize the melody of the saxophone. One of them said that the melody was good compared to the original excerpt but not with the ground truth, another commented that the melody was even better than the proposed ground truth and finally one subject explained her doubts about how to grade this excerpt.

Type	Glissando	Far	Near	Other instrument
Median	4	3	3	0,5
Average	4,1	2,8	3	0.9
STD	0,8	0,9	1,2	1,3
Mode	4	3	4	0

Table 3.5: Results for the excerpts containing errors with other pitches.

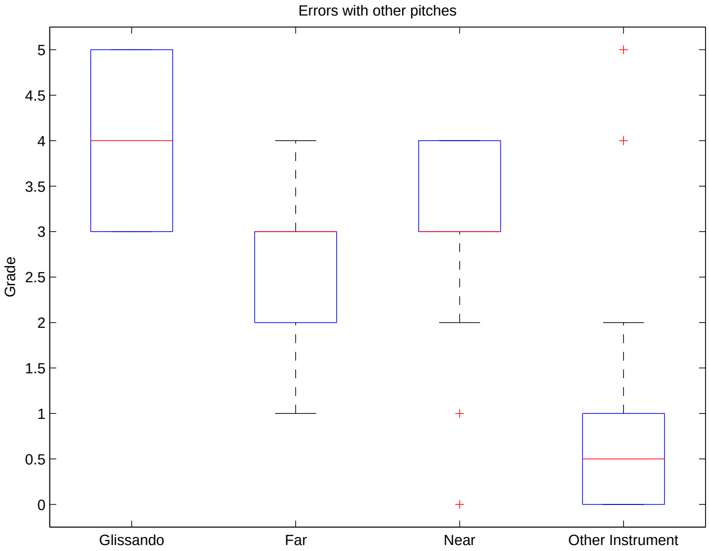


Figure 3.15: User ratings for excerpts containing different pitch errors.

Next, the results for the excerpts selected to demonstrate the hypothesis 5 are shown in Table 3.6 and in Fig. 3.16. The two excerpts have the same Overall Accuracy (85%) and we can observe that the medians of these excerpts are equal. Furthermore, after reading the comments, we can affirm that 2 subjects were able to distinguish the octave errors while 5 were not able. So, the hypothesis 5 is false.

Type	Octave error	Error with other pitch
Median	2	2
Average	2,2	2,2
STD	1,2	1
Mode	2	3

Table 3.6: Octave error results.

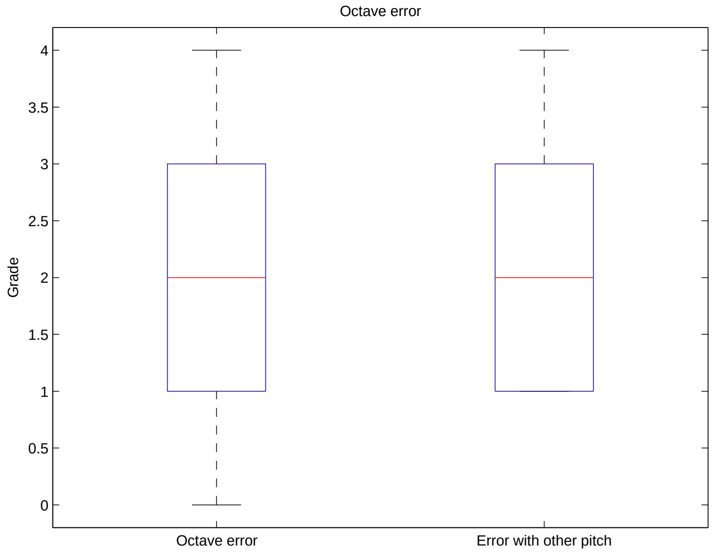


Figure 3.16: User ratings for two excerpts containing an octave error and an error with other pitch respectively.

Finally, the results of the excerpts for testing the hypothesis 7 about the vibrato are shown in Table 3.7 and in Fig. 3.17. We can appreciate that there is one point of difference between the medians of both excerpts. Furthermore, there were six comments complaining about the vibrato for the excerpt with a bad reconstruction. So, the hypothesis 7 is true.

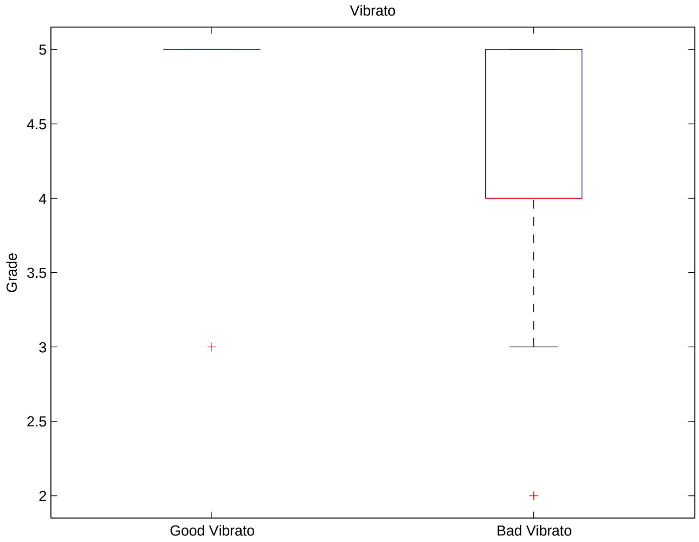


Figure 3.17: User ratings for two excerpts with a good and a bad reconstruction of the vibrato respectively.

Vibrato	Good	Bad
Median	5	4
Average	4,8	4,2
STD	0,5	1
Mode	5	5

Table 3.7: Vibrato results.

3.4.3 Time Continuity

The results for the excerpts selected to test the hypothesis 8 are shown in Table 3.8 and in Fig. 3.18. The two excerpts have the same Overall Accuracy (90%) and we can observe that there is half a point of difference between the excerpt containing one error and the excerpt containing two errors. So, the hypothesis is true at least for the pitch errors. Moreover, the subjects explained in their comments that they have noticed that in the first case there is only one error while in the second case there are two errors.

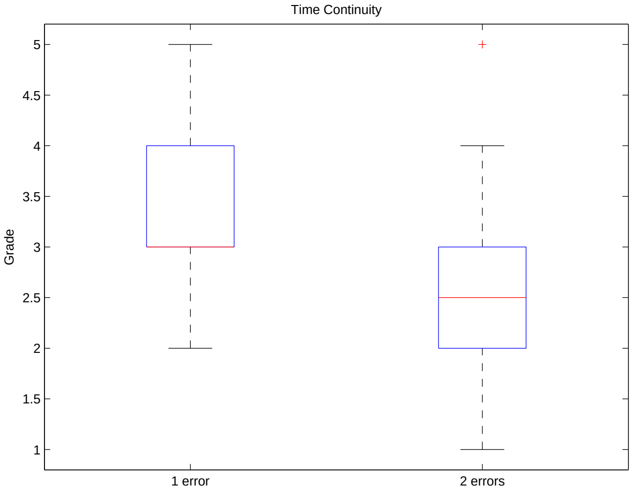


Figure 3.18: User ratings for two excerpts with one and two grouped pitch errors.

Errors	1	2
Median	3	2,5
Average	3,3	2,5
STD	0,8	1
Mode	3	3

Table 3.8: Time continuity results.

3.4.4 Recap

This section presents in the Table 3.9 a brief summary of the results of the survey. We confirmed 5 of the 8 hypotheses proposed which are related to the importance of the melodic context, the vibrato reconstruction and the time continuity or the distribution of the errors.

Hypothesis	Error	Topic	State
1	False alarms	Pitch and time position relative to the melodic context.	Confirmed with the case of lengthening a note vs. adding random noise.
2	Missing frames	Pitch and time position relative to the melodic context.	Confirmed with the case of shortening a note vs. missing a full note.
3	Pitch errors	Pitch and time position relative to the melodic context.	Confirmed with the case of adding a glissando vs. generic error.
4	Pitch errors	Frequency distance.	There is insufficient evidence to confirm it.
5	Octave errors	Octave errors vs. Errors with other pitches.	It is false because most of the users cannot distinguish between octave errors and errors with other pitches.
6	Pitch errors	Confusion with other instruments different from the predominant one.	There is insufficient evidence to confirm it, although some users identify the solo of a saxophone as the predominant instrument instead of the voice.
7	None	Reconstruction of the vibrato.	Confirmed with the case of a good reconstruction vs. a bad reconstruction.
8	All	Time continuity and grouped errors.	Confirmed in the case of pitch errors.

Table 3.9: Summary of the results.

Chapter 4

CONCLUSION

After carrying out the survey, this chapter sums up the main conclusions achieved. Furthermore, it also explains the future perspectives which include the model proposed to build a new metric closer to the perception of the users called Accuracy by Users.

4.1 Discussion and ideas for metric enhancements

Although modelling how the users perceive the melodies properly requires more surveys, in some cases including their opinion in a new metric should be relatively easy. We propose a modification of the current overall accuracy in which we will include corrections taking into account the different aspects that affect the perception of the users. The current overall accuracy is the proportion of all frames correctly estimated by the algorithm, which means that both the pitch estimation and the voicing detection are right. But during the analysis of the results of the survey, we could check that some mistaken frames go unnoticed by the users or have different impact on the perceived quality depending on the kind of error. Moreover, we also check that some correct frames including a bad reconstruction of the vibrato caused a bad effect on the quality. So, these frames should be taken into account in order to include corrections in the overall accuracy related to the hypotheses that were proved during the survey, achieving this way a new metric that represents better the opinion of the users. We call this new metric Accuracy by Users:

$$Acc_{users} = Acc_{ov} + \frac{1}{L}\sum_{t} Correction_t$$

$$Acc_{users} = \frac{1}{L}\sum_{t} v_t^* T[M(f_t) - M(f_t^*)] + \overline{v}_t^* \overline{v}_t + Correction_t$$

Where **Correction** is a vector that can be defined for the frame t as:

- If $v_t^* T[M(f_t) - M(f_t^*)] + \overline{v}_t^* \overline{v}_t = 1$ (the frame is considered correct following the criteria of the overall accuracy) and the frame belongs to a fragment containing a vibrato with a bad reconstruction, the correction will be a negative value: $-1 \leq Correction_t \leq 0$. In this case, -1 means that the bad reconstruction is as bad as an incorrect frame while 0 means that it goes unnoticed. The intermediate values will depend on the answers of the users to different cases, including different depth values or reconstruction errors.

- If $v_t^* T[M(f_t) - M(f_t^*)] + \overline{v}_t^* \overline{v}_t = 0$ (the frame is considered incorrect) and it belongs to a special kind of error, including lengthening or shortening a note or adding a glissando, the correction will be a positive value: $0 \leq Correction_t \leq 1$. In this case, 0 means that the error goes unnoticed by the users while 1 means that the special error is as bad as a generic mistake. The intermediate values will depend on how the users perceive different duration of the errors, being correction higher for shorter errors.
- Otherwise, $Correction_t = 0$, including correct frames without vibrato or with a good reconstruction and incorrect frames that do not belong to a special kind of error.

When implementing this new metric we have focused first on the issue of lengthening or shortening a note, because it is the error with more variability of cases in our dataset and with more excerpts included in the survey. In the survey we considered absolute duration of the errors in frames or seconds without taking into account the duration of the lengthened/shortened note or other notes in the melodic context. However, we should carry out another survey with more excerpts and considering the relative duration of the errors, because the perception of the users about the errors will depend both on the duration of the lengthened/shortened note and on the duration of the error. For example, it is clear that it is not the same adding 1 sec to a note of 5 sec than adding 1 sec to a note of 0.2 sec. Moreover, *a priori* it is not the same moving the onset than moving the offset. So, we will need to build a different model for each possible case:

- Lengthening a note moving forwards offset.
- Lengthening a note moving backwards the onset.
- Shortening a note moving backwards the offset.
- Shortening a note moving forwards the onset.

Moreover, we should not get confused by the notation used in this master thesis to denote the case of lengthening or shortening a note, because we are evaluating the melody extraction task by frames, not the transcription task. Therefore, we should be able to evaluate these issues without passing through the transcription, because that task has its own evaluation metrics which work fine. The idea in this case is to take into account the opinion of the users when deciding if a frame is correct or not, but preserving the main concept of a evaluation by frames. So, in the case of lengthening a note we look for groups of consecutive false alarms with the same pitch than a previous or following consecutive group of correctly estimated voiced frames. In the case of shortening a note, we look for groups of consecutive missing frames missing the same pitch than a previous or following consecutive group of correctly estimated voiced frames. Looking this way for the four cases cited before.

After that, we define the Normalized Error Duration (NED) as the ratio between the duration of the error and the duration of the group of correctly estimated voiced frames with the same pitch in seconds. Taking this into account we should carry out a survey to model the perception of the users as a function of the NED, which probably will have a shape similar to the one showed at Fig. 4.1, where:

- If $NED < L$, the mistake will go unnoticed by the user and $Correction_t = 1$ for the incorrect frames involved in the error.
- If $L < NED < H$, the mistake is noticed and the impact on the perceived quality will be increasing with the NED. So, $Correction_t$ decreases with increasing NED (maybe linearly).
- If $NED > H$, the mistake is noticed by the users in a way that the perceived quality of a special error is equal to that of a generic error. Therefore, $Correction_t = 0$ for the frames belonging to the grouped error.

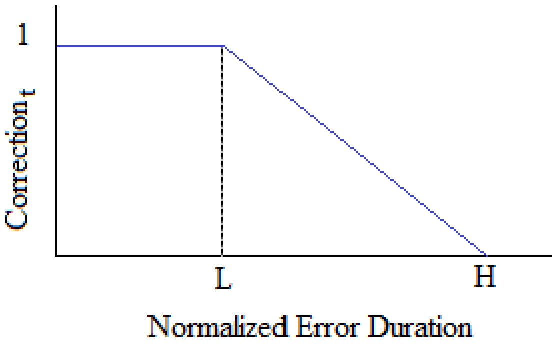


Figure 4.1: Correction vs. Normalized Error Duration.

Moreover, the values of L and H probably will be different for each of the cases cited before. Although, we have not yet conducted new surveys to find those values, we have evaluate some excerpts belonging to the previous survey using values for L and H set by ourselves. However, this evaluation is not rigorous because these values represent only our opinion without asking people. In any case, some preliminary results show that it is possible to achieve a metric closer to how the users perceive the melodies.

Furthermore, with the proposed model it is possible to achieve accuracies closer to 100% when the users do not notice the mistakes made by the algorithms, offering the possibility of breaking the glass ceiling by involving the users. However, at this point it is very important to take into account that the perception of the users could be different depending on the specific application of the melody extracted. For example, the requirements of a query by humming system may be different from those of a transcription system, and the methodology employed until now to model the users perception may not be appropriated for some applications. So, another future line is to carry out other surveys oriented to specific applications. Nevertheless, the model proposed for the new metric based on corrections in the overall accuracy is flexible and easily adaptable to the specific needs of each application, letting us to include the results of new surveys in a scalable way.

Finally, it is also important to mention that the hypothesis relative to the time continuity has not been included yet in the Accuracy by Users. A direct way of including it is to multiply it by 1/G where G is the number of grouped errors present in the excerpt. However, we also need to carry out another survey with more variety of cases, because we know that the quality decreases with the number of grouped errors but we have not enough information to determine in which proportion or way.

4.2 Conclusion

Taking into account the results of the survey we can conclude that the current metrics are not properly related to how the users perceive the extracted melodies. First, the only one that evaluates the overall system is the overall accuracy. So, it is the only one with which we can compare the results of the survey, because the output of the intermediate steps does not reach directly the user.

We could check that a higher overall accuracy does not always involve a higher grade by the users in the perceptual evaluation, because the actual per frame evaluation is not necessarily related to the opinion of the users. This is due to the fact that we have disregarded the pitch relation between the frames and the temporal distribution of the errors, but these issues affect dramatically the perception of the users.

In Chapter 3, we proposed eight hypotheses and we have been able to demonstrate five of them. First, we proved that lengthening or shortening a note has less impact in the perceived quality than adding a random noise or missing a full note. Then, we demonstrated that including a glissando affects less the quality of the melody than a generic pitch error. Next, we check the importance of a good vibrato reconstruction. Finally, we also checked the importance of the time continuity and the distribution of errors, at least for pitch errors.

Finally, we also noticed that some users perceive as ground truth the melody played by other instrument different from the one proposed. That is, we disagree in which was the predominant melody. These could be taken into account in the future by conducting user surveys when establishing the ground truth. Another option is to include two pondered ground truths with the melody played by the two more predominant instruments, such as in other tasks in which the perception is a personal issue (e.g. tempo estimation).

4.3 Future perspectives

This research has been a first step to realize the importance of the perceptual evaluation in melody extraction. However, further experiments should be carried out to reach more conclusions and propose a good new evaluation metric. All the cases proposed in this survey belongs to real cases of our dataset. However, if we extend the dataset or if we synthesise other outputs that we could make up, the possibilities will improve considerably because we will be able to include more variety of cases in a more controlled context. For instance, we know that the reconstruction of the vibrato affects the perception of the user, but how deep should it be to be noticed by the users? Another question could be how long should be a glissando to have a small penalization by the users. These are questions that need a bigger survey with a high number of excerpts in which almost nobody will be willing to participate for free.

Furthermore, we should go deeper into the perception of the melodies to understand better what is happening. For instance, when lengthening or shortening a note, how different is the effect of moving the onset versus moving the offset? Moreover, the perception in this case may be a relative issue. That is, instead of considering an absolute value for the duration of the errors, maybe we should select durations relative to the total length of the note, or even take into account the duration of all the notes in the melodic context and involve tempo issues. So, definitely it will be a hard, expensive and long task to properly characterize the melodic perception of the users, because we can not find the proper answers in the literature and we should carry out at least one survey for each question that we want to respond.

Nevertheless, the evaluation process is very important for improving the algorithms. Moreover, since the ultimate goal of the evaluation is to characterize the usage experience of the users, it is worth to researching on new evaluation metrics closer to the perception of the users. Furthermore, as we said before depending on the application of the melody extraction task some errors may be more relevant than others (e.g. transcription vs. music similarity tasks), therefore the new metrics should be aware of the context and be adapted for specific future applications of the extracted melody if necessary.

Bibliography

Arora, V. and Behera, L. (2013). On-line melody extraction from polyphonic audio using harmonic cluster tracking. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(3):520–530.

Biehl, J. T., Adamczyk, P. D., and Bailey, B. P. (2006). Djogger: a mobile dynamic music device. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, pages 556–561. ACM.

Bogdanov, D. and Herrera, P. (2011). How much metadata do we need in music recommendation? a subjective evaluation using preference sets. In *ISMIR*, pages 97–102.

Cancela, P. (2008). Tracking melody in polyphonic audio. mirex 2008. *Proc. of Music Information Retrieval Evaluation eXchange*.

Charpentier, F. (1986). Pitch detection using the short-term phase spectrum. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, volume 11, pages 113–116. IEEE.

Downie, J. S., West, K., Ehmann, A., Vincent, E., et al. (2005). The 2005 music information retrieval evaluation exchange (mirex 2005): Preliminary overview. In *6th Int. Conf. on Music Information Retrieval (ISMIR)*, pages 320–323.

Dressler, K. (2005). Extraction of the melody pitch contour from polyphonic audio. In *Proc. 6th International Conference on Music Information Retrieval*, volume 108.

Dressler, K. (2006). Sinusoidal extraction using an efficient implementation of a multi-resolution fft. In *Proc. of 9th Int. Conf. on Digital Audio Effects (DAFx-06)*, pages 247–252.

Dressler, K. (2009). Audio melody extraction for mirex 2009. *5th Music Information Retrieval Evaluation eXchange (MIREX)*.

Dressler, K. (2011). An auditory streaming approach for melody extraction from polyphonic music. In *ISMIR*, pages 19–24.

Durrieu, J.-L., Richard, G., David, B., and Févotte, C. (2010). Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):564–575.

Firan, C. S., Nejdl, W., and Paiu, R. (2007). The benefit of using tag-based profiles. In *Web Conference, 2007. LA-WEB 2007. Latin American*, pages 32–41. IEEE.

Fuentes, B., Liutkus, A., Badeau, R., and Richard, G. (2012). Probabilistic model for main melody extraction using constant-q transform. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5357–5360. IEEE.

Goto, M. (2004). A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4):311–329.

Hartmann, W. M. (1997). *Signals, sound, and sensation*. Springer.

Hsu, C.-L. and Jang, J.-S. R. (2010). Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion. In *ISMIR*, pages 525–530.

Hu, X. and Liu, J. (2010). Evaluation of music information retrieval: Towards a user-centered approach. In *Proceedings of the 4th Workshop on Human-Computer Interaction and Information Retrieval (HCIR). New Brunswick, NJ,(August 2010)*.

Kaminskas, M. and Ricci, F. (2011). Location-adapted music recommendation using tags. In *User Modeling, Adaption and Personalization*, pages 183–194. Springer.

Klapuri, A. (2004). *Signal processing methods for the automatic transcription of music*. Tampere University of Technology Finland.

Knees, P., Pohle, T., Schedl, M., and Widmer, G. (2007). A music search engine built upon audio-based and web-based similarity measures. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 447–454. ACM.

Lee, J. S. and Lee, J. C. (2007). Context awareness by case-based reasoning in a music recommendation system. In *Ubiquitous Computing Systems*, pages 45–58. Springer.

Liu, H., Hu, J., and Rauterberg, M. (2009). Music playlist recommendation based on user heartbeat and music preference. In *Computer Technology and Development, 2009. ICCTD'09. International Conference on*, volume 1, pages 545–549. IEEE.

Marolt, M. (2005). Audio melody extraction based on timbral similarity of melodic fragments. volume II, pages 1288–1291. cited By (since 1996)1.

Moens, B., van Noorden, L., and Leman, M. (2010). D-jogger: Syncing music with walking. In *7th Sound and Music Computing Conference*, pages 451–456. Universidad Pompeu Fabra.

Paiva, R. P., Mendes, T., and Cardoso, A. (2006). Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness. *Computer Music Journal*, 30(4):80–98.

Pauws, S. and Eggen, B. (2002). Pats: Realization and user evaluation of an automatic playlist generator. In *ISMIR*.

Pauws, S. and van de Wijdeven, S. (2005). User evaluation of a new interactive playlist generation concept. In *ISMIR*, pages 638–643.

Poliner, G. E. and Ellis, D. P. (2005). A classification approach to melody transcription. In *ISMIR 2005: 6th International Conference on Music Information Retrieval: Proceedings: Variation 2: Queen Mary, University of London & Goldsmiths College, University of London, 11-15 September, 2005*, pages 161–166. Queen Mary, University of London.

Poliner, G. E., Ellis, D. P., Ehmann, A. F., Gómez, E., Streich, S., and Ong, B. (2007). Melody transcription from music audio: Approaches and evaluation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(4):1247–1256.

Rafii, Z. and Pardo, B. (2013). Repeating pattern extraction technique (repet): A simple method for music/voice separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(1):73–84.

Rao, V. and Rao, P. (2010). Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(8):2145–2154.

Ryynänen, M. P. and Klapuri, A. P. (2008). Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86.

Salamon, J. (2012). Ame annotation initiative.

Salamon, J. and Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(6):1759–1770.

Salamon, J., Gomez, E., Ellis, D., and Richard, G. (2014). Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *Signal Processing Magazine, IEEE*, 31(2):118–134.

Salamon, J. and Urbano, J. (2012). Current challenges in the evaluation of predominant melody extraction algorithms. In *ISMIR*, pages 289–294.

Schedl, M., Flexer, A., and Urbano, J. (2013). The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3):523–539.

Slaney, M. and Lyon, R. F. (1993). On the importance of time-a temporal representation of sound. *Visual representations of speech signals*, pages 95–116.

Sutton, C., Vincent, E., Plumbley, M. D., Bello, J. P., et al. (2006). Transcription of vocal melodies using voice characteristics and algorithm fusion. In *2006 Music Information Retrieval Evaluation eXchange (MIREX)*.

Tachibana, H., Ono, T., Ono, N., and Sagayama, S. (2010). Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 425–428. IEEE.

Urbano, J. (2013). *Evaluation in Audio Music Similarity*. PhD thesis, University Carlos III of Madrid.

Urbano, J., Downie, J. S., Mcfee, B., and Schedl, M. (2012). How Significant is Statistically Significant? The case of Audio Music Similarity and Retrieval. In *International Society for Music Information Retrieval Conference*, pages 181–186.

Urbano, J., Schedl, M., and Serra, X. (2013). Evaluation in Music Information Retrieval. *Journal of Intelligent Information Systems*, 41(3):345–369.

Vignoli, F. and Pauws, S. (2005). A music retrieval system based on user driven similarity and its evaluation. In *ISMIR*, pages 272–279. Citeseer.

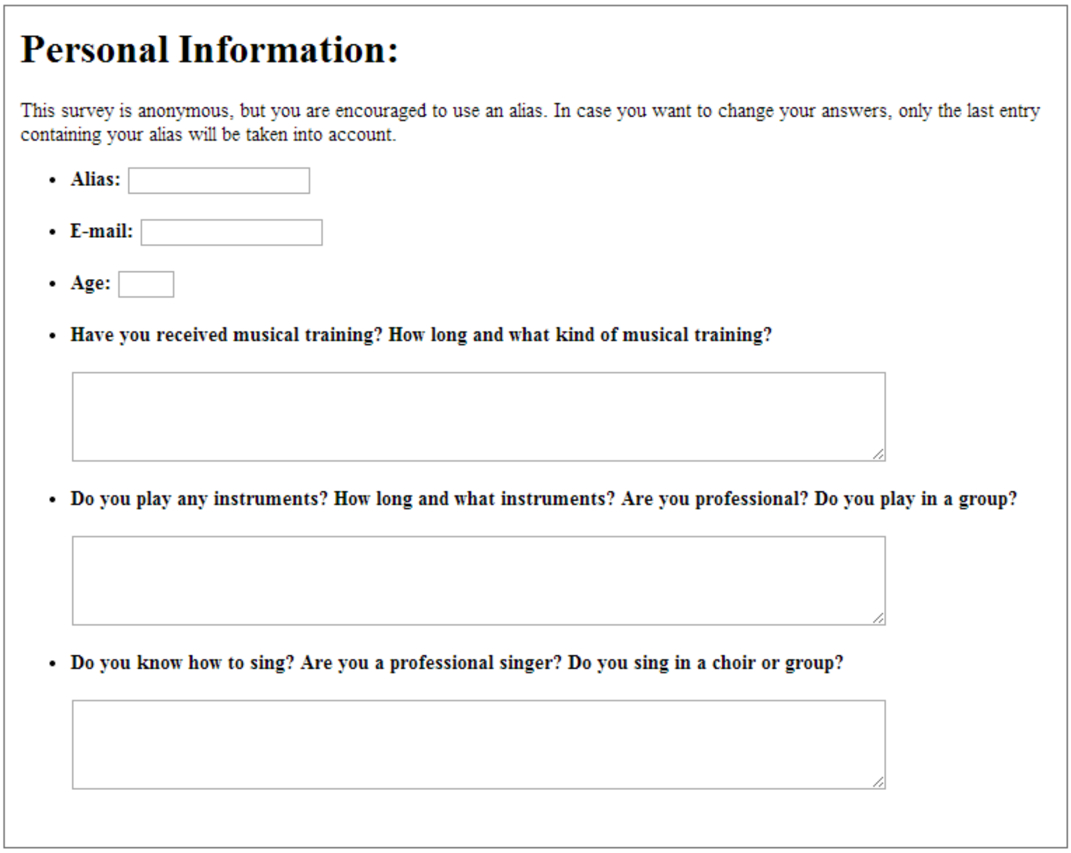
Xue, G.-R., Han, J., Yu, Y., and Yang, Q. (2009). User language model for collaborative personalized search. *ACM Transactions on Information Systems (TOIS)*, 27(2):11.

Yeh, T.-C., Wu, M.-J., Jang, J., Chang, W.-L., and Liao, I.-B. (2012). A hybrid approach to singing pitch extraction based on trend estimation and hidden markov models. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 457–460. IEEE.

Appendices

Appendix A

FORM

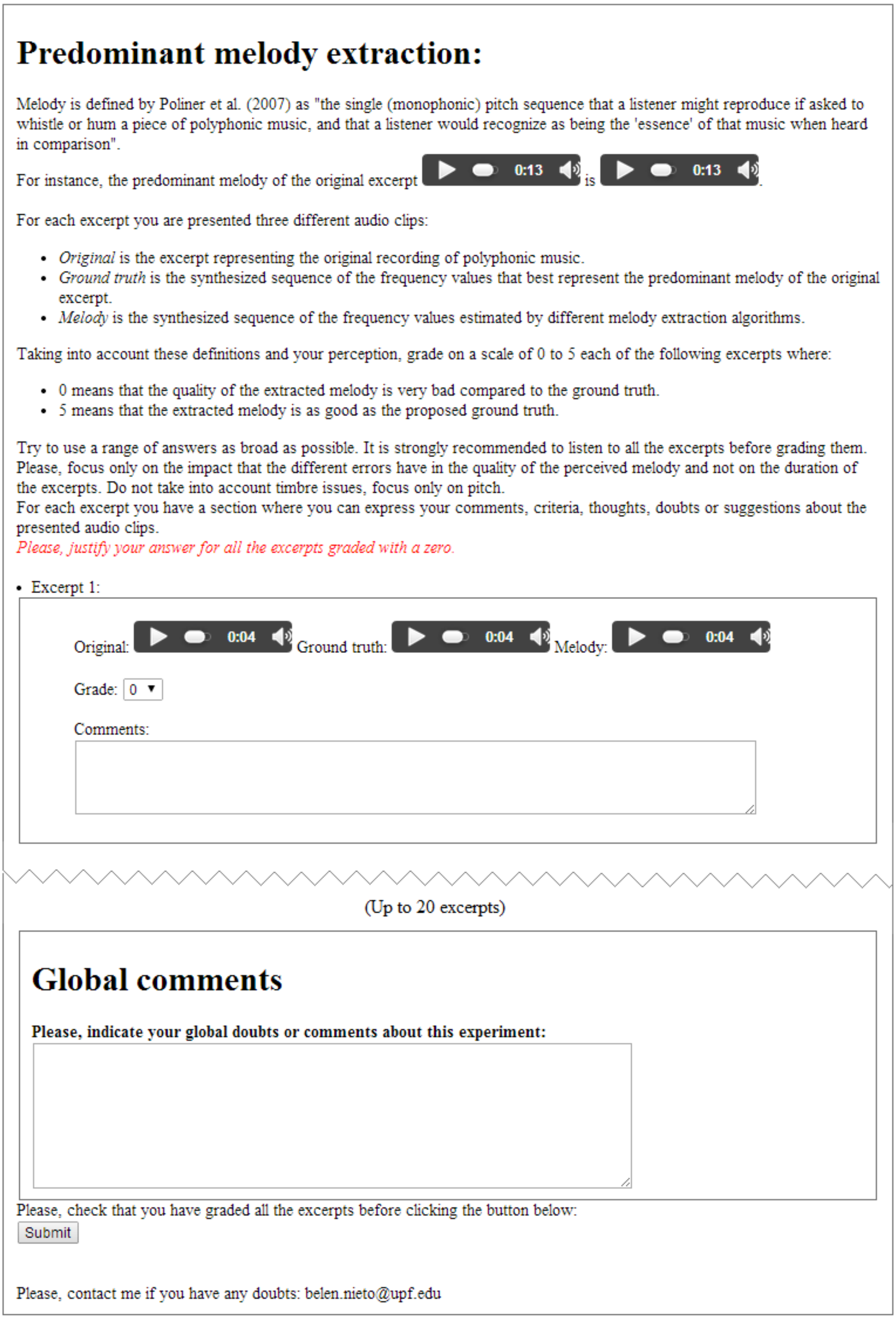


Personal Information:

This survey is anonymous, but you are encouraged to use an alias. In case you want to change your answers, only the last entry containing your alias will be taken into account.

- **Alias:** ________
- **E-mail:** ________
- **Age:** ____
- **Have you received musical training? How long and what kind of musical training?**
- **Do you play any instruments? How long and what instruments? Are you professional? Do you play in a group?**
- **Do you know how to sing? Are you a professional singer? Do you sing in a choir or group?**

Figure A.1: Form.



Predominant melody extraction:

Melody is defined by Poliner et al. (2007) as "the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music, and that a listener would recognize as being the 'essence' of that music when heard in comparison".

For instance, the predominant melody of the original excerpt ▶ 0:13 is ▶ 0:13.

For each excerpt you are presented three different audio clips:

- *Original* is the excerpt representing the original recording of polyphonic music.
- *Ground truth* is the synthesized sequence of the frequency values that best represent the predominant melody of the original excerpt.
- *Melody* is the synthesized sequence of the frequency values estimated by different melody extraction algorithms.

Taking into account these definitions and your perception, grade on a scale of 0 to 5 each of the following excerpts where:

- 0 means that the quality of the extracted melody is very bad compared to the ground truth.
- 5 means that the extracted melody is as good as the proposed ground truth.

Try to use a range of answers as broad as possible. It is strongly recommended to listen to all the excerpts before grading them. Please, focus only on the impact that the different errors have in the quality of the perceived melody and not on the duration of the excerpts. Do not take into account timbre issues, focus only on pitch.
For each excerpt you have a section where you can express your comments, criteria, thoughts, doubts or suggestions about the presented audio clips.
Please, justify your answer for all the excerpts graded with a zero.

- Excerpt 1:

Original: ▶ 0:04 Ground truth: ▶ 0:04 Melody: ▶ 0:04

Grade: 0 ▼

Comments:

(Up to 20 excerpts)

Global comments

Please, indicate your global doubts or comments about this experiment:

Please, check that you have graded all the excerpts before clicking the button below:

Please, contact me if you have any doubts: belen.nieto@upf.edu

Figure A.2: Melody extraction form.

Appendix B

SURVEY RESULTS

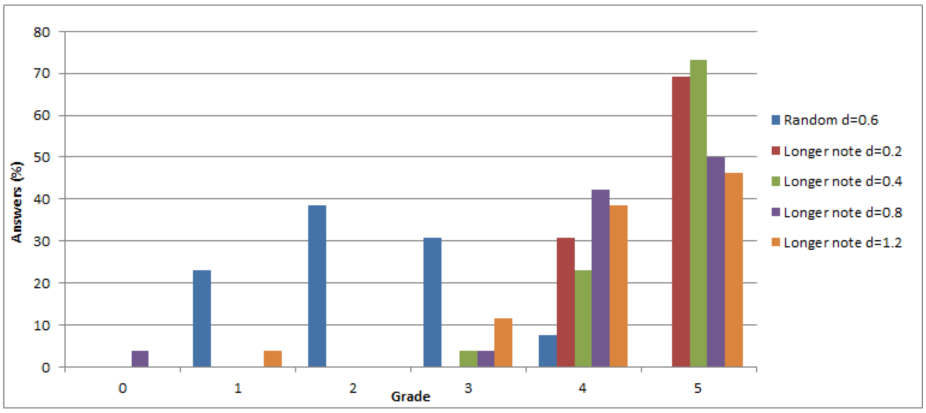


Figure B.1: User ratings for the excerpts containing false alarms.

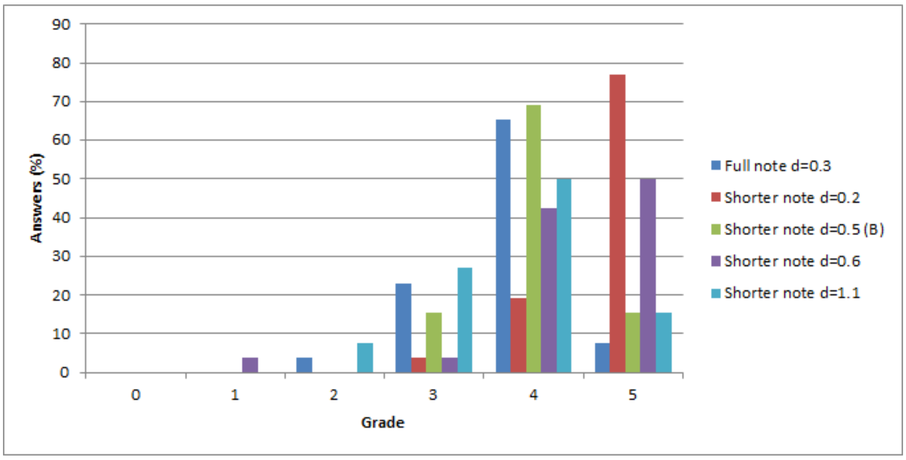


Figure B.2: User ratings for the excerpts containing missing frames.

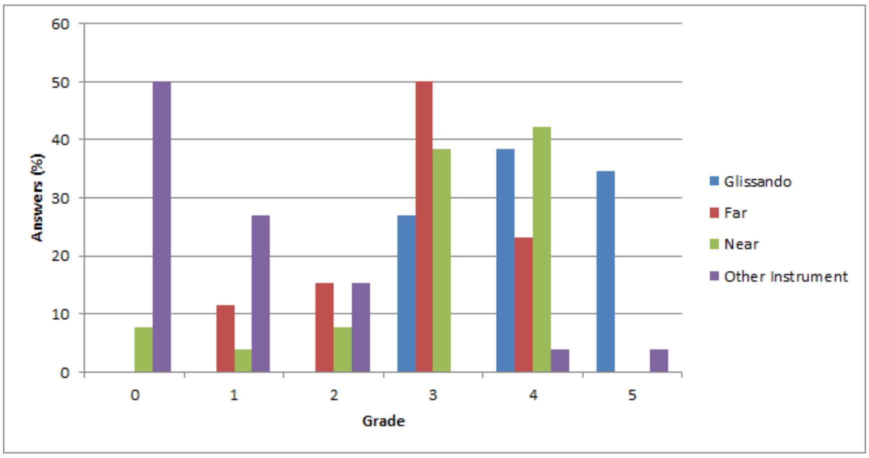


Figure B.3: User ratings for excerpts containing different pitch errors.

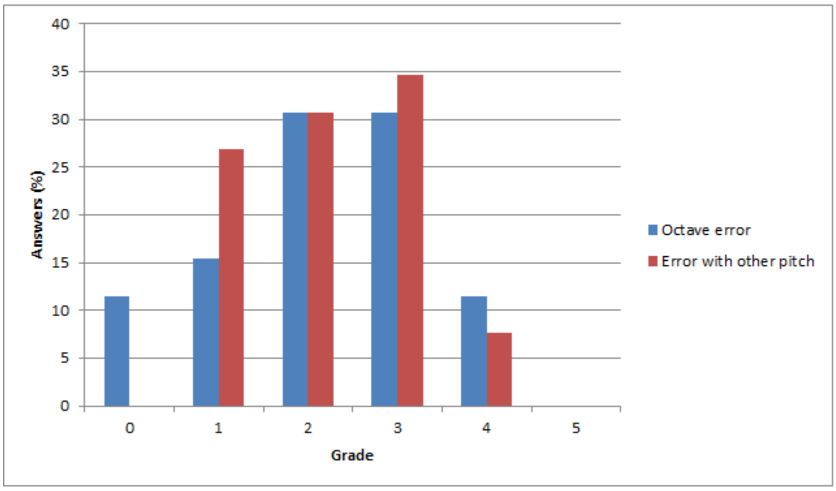


Figure B.4: User ratings for two excerpts containing an octave error and an error with other pitch respectively.

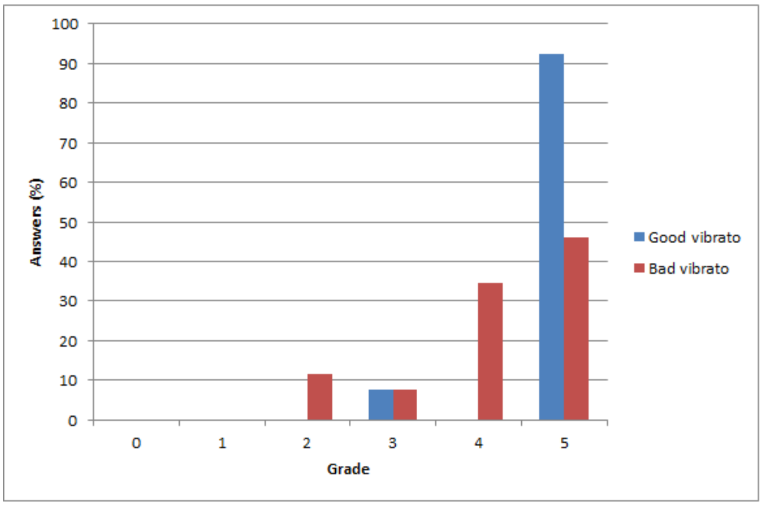


Figure B.5: User ratings for two excerpts with a good and a bad reconstruction of the vibrato respectively.

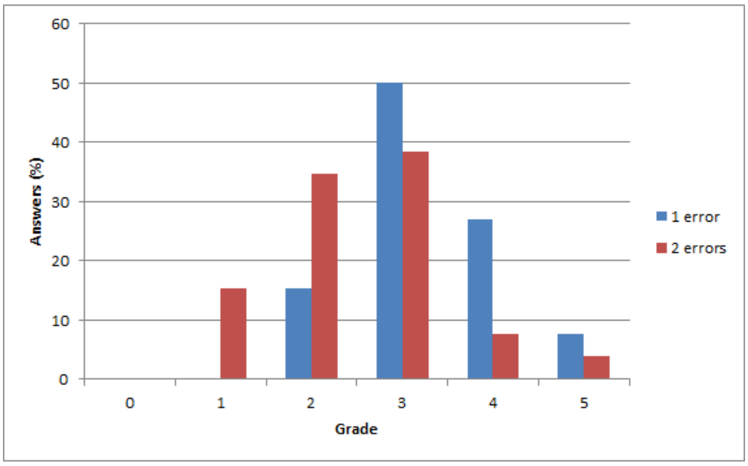


Figure B.6: User ratings for two excerpts with one and two grouped pitch errors respectively.