Study of reverberation robust pitch estimators for the singing voice

Hèctor Parra Rodríguez

MASTER THESIS UPF 2013

Master in Sound and Music Computing

Master thesis supervisor: Dr. Jordi Janer

Host: Music Technology Group

Department of Information and Communication Technologies

Universitat Pompeu Fabra, Barcelona

upf. Universitat Pompeu Fabra *Barcelona*

Acknowledgments

First of all, I want to thanks Xavier Serra for giving me the opportunity to course this master. I remember the first days of class he said this master would either be another line on our CV or change our lives, and I can now say I belong to the last option. He had the courage to create a beautiful research group like the MTG in Barcelona, and without it I would have never been able to discover the thrilling world of music computing. I want to thank him for teaching me exactly what I came for, for understanding and helping me every time I complained about any organizational issue, and for giving me the opportunity to work with him.

I am grateful to Jordi Janer for supervising this master thesis. He believed in me and clarified my mind in the tough moments. I had lots of long but enriching discussions with him without which it would have never been possible to carry out this thesis. Many exciting ideas had finally stayed as a draft but I hope some day we can resume them.

I extend my gratitude to Emilia Gómez for providing me with lots of materials; algorithms, good references but more importantly to the priceless classes she gave me.

I want to thank Justin Salomon to be always willing to assist me with MELODIA and giving me valuable pieces of advice.

I am obliged to Emanuël A.P. Habets, DeLiang Wang and Donald S. Williamson for replying my emails and helping me with their de-reverberation algorithms.

I want to mention Vanessa Jimenez, for helping me with the endless paper work and my last minute changing of subjects.

I want to thank Sonia Espí and Jordi Janer again for making it possible to me to give a talk in Sonar 2013 as well as having a great collaboration with the musical band Za!.

I am grateful to Alba Rosado for handling all the details that allowed me to participate in many different music hack events, especially when problems came up abroad.

To all my classmates, who brought the fun but also lots of wisdom and without whom I would have never made this master through.

Last but not the least, I am grateful to my family, especially Inma Rodriguez and Carmen de Terán, to give me support when the days were blue and encourage me to pursue the things that I love.

Abstract

Making machines understand us has been a challenging issue in the later years. Although reverberation is an omnipresent phenomenon in our daily lives, computers are still not prepared to handle it correctly. A study to help machines overcome reverberation when estimating fundamental frequency is presented.

The study focuses on the singing voice since it is the form of human expression with more complex fundamental frequency contours. There have been selected four fundamental frequency estimation algorithms (YIN, TWM, SAC, MELODIA) common for this task in dry conditions. The study evaluates them following the MIREX Audio Melody Extraction evaluation criteria.

First, the algorithms are evaluated in dry conditions and different reverberant conditions. It is shown how an increasing reverberation time supposes an increasing loss in accuracy for all algorithms. Besides, MELODIA exhibits a special robustness compared to its competitors.

Then, we try to improve fundamental frequency estimators' performance using different de-reverberation methods (NML, NMF, ITD) as preprocessors. Only NML succeeds in such a task for all algorithms except MELODIA, which keeps performing the best. Anyway, it demonstrates that de-reverberation methods can be used to improve fundamental frequency estimators' results in reverberant conditions.

Finally, the insights of the study results are analyzed. In order to exemplify how the results of this study can be used to improve algorithms' accuracy, a proof-of-concept algorithm (MIX) is presented. MIX combines MELODIA with SAC and NML de-reverberation. It has a general improvement in accuracy of 2% in reverberant conditions and, in addition, it performs as good as the best algorithms in dry conditions: 91% overall accuracy.

Contents

1 Introduction

This chapter introduces the problem that will be faced and the motivations that justify its interest of study. Then our objectives are presented and, finally, a brief overview of this dissertation is outlined.

1.1 Motivation

Since the most remotely times, the singing voice has been our most accessible instrument for the simple fact it comes incorporated in our body. Consequently, it has always been a powerful tool of expression; be it for praying to some God to make it rain or to remember a friend about that tune you love. But it is not just this universality that makes it special. It is the ability of creating a channel of communication that goes beyond the rationality of speech and becomes closer to the expression of feelings what makes it unique. Moreover, it is the only immediate tool we posses for expressing melodic content.

Since the last three decades, there has been a significant effort in the audio and music signal-processing field to make computers able to capture and understand some of this expressiveness, especially for the melodic content. This dissertation pretends to contribute in this same direction.

Nowadays, it is possible for computers to “listen” to a human singing and be able to transcribe more or less accurately the score of the sung melody. This accuracy depends on many factors as the style of singing, the range of the voice, the microphone used, interfering noises, environmental effects, etc. We focus on a particular environmental effect: reverberation. Reverberation has been shown to be an injurious effect for computers when trying to extract the pitch of a sung melody and, more importantly, they suffer significant accuracy degradation for reverberations that humans handle easily.

Lately, thanks to the technology expansion, we are living the mobile device era. There have become more common scenarios where machines have to deal with the singing voice with a significant reverberation amount. For example, singing to mobile phones to identify a particular song we cannot recall. Or singing in Karaokes that rank our performance. Even some standard-bearer artists feed their voices in special devices to get them transformed in new creative ways. Indeed, my personal motivation for this study came from the idea of analyzing the singing voice to control artificial synthesizers.

In the past few years, there have appeared multiple approaches for removing the reverberation effect from speech signals, but a very few have tackled the problems it causes to fundamental frequency estimation. Several degradation problems have been observed: non-existing pitch is detected on silences after sounds, unvoiced consonants disappear being surpassed by vocals or pitch detection is just wrong because old sounds keep still sounding. This dissertation studies these problems.

1.2 Goals

The main goal of this thesis is to study the negative effects that reverberation introduces when trying to estimate the fundamental frequency of a singing voice signal and, having this knowledge, point out the direction for removing these effects. Some questions we want to answer are:

- What are the (negative) effects that reverberation produces on fundamental frequency estimation algorithms?
- From the existing fundamental frequencies estimators, which one is the most robust to reverberation effects? Why? Can we take advantage of some robust strategy it is using?
- Can we achieve better results on fundamental frequency estimation applying current reverberation reduction algorithms?
- If so, which is the reverberation reduction algorithm that improves more the results? Why? Can we take advantage of some robust strategy it is using?
- Do reverberation reduction algorithms improve more the results of some fundamental frequency estimators than others? Why?
- Which are the best evaluation methods for getting the answer to the precedent questions?

Once these questions have been answered, it will be proposed an algorithm for fundamental frequency estimation, created from the extracted ideas, and finally evaluated and compared to the previous ones.

1.3 Thesis overview

This dissertation is organized as follows. Section 2 presents an overview on fundamental frequency estimators; how they operate, their parts, how their different algorithms operate and a final special mention for those estimators dedicated to the singing voice or reverberated signals. Section 3 describes the reverberation topic. First, acoustic room principles are described. Second, speech reverberation reduction methods are outlined. And finally, the effects that reverberation produces on the singing voice are explained. Section 4 explains the insights of the study carried out in this thesis; methodology, evaluation methods, algorithms used and datasets selected. Section 5 analyses the results obtained from the study, presenting partial conclusions that are then summarized as final conclusions in Section 6. Section 7 points out the next steps to be carried to further research in the direction of this study. Finally, an appendix with the numerical results is attached in Section 8.

2 Fundamental frequency estimation

This chapter introduces the concept of fundamental frequency. Then, it presents an overview (based on [1]) of the existing methods for fundamental frequency estimation and describes their process steps. Finally, it focuses on the current methods that are of interest for this study, i.e. those involving the singing voice and reverberated signals.

2.1 Introduction

The fundamental frequency (f_0), also simply referred to as fundamental, is defined as the lowest frequency of a periodic waveform. In terms of a superposition of sinusoids, the fundamental frequency is the lowest frequency sinusoidal in the sum. All sinusoidal and many non-sinusoidal waveforms are periodic, which is to say they repeat exactly over time. A single period is thus the smallest repeating unit of a signal, and one period describes the signal completely. We can show a waveform is periodic by finding some period T_0 (the fundamental period) for which the following equation is true:

$$x(t) = x(t + nT_0) \tag{2.1}$$

where $x(t)$ is the function of the waveform, t indicates time and $n \in \mathbb{N}$. Then, the relation with the fundamental frequency is $f_0 = 1/T_0$ (see Figure 2.1). It is important to note that while f_0 and pitch are different concepts (the former being a physical measurement and the latter a perceptual one) they are commonly used indistinguishably in the literature.

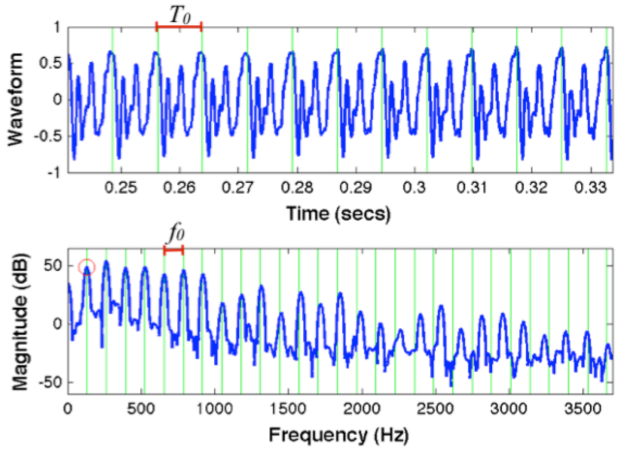

Figure 2.1 Fundamental period (above) in time domain and fundamental frequency (below) in frequency domain

All fundamental frequency algorithms give a measure corresponding to a portion of the signal (analysis frame). The fundamental frequency estimation process can be subdivided into three main steps (according to [2]) that are passed through successively: the pre-processor, the basic extractor, and the post-processor (see Figure 2.2). The basic extractor performs the main task of measurement: it converts the input signal into a series of fundamental frequency estimates. The main task of the pre-processor is to reduce the amount of data in order to facilitate the fundamental frequency extraction. Finally, the postprocessor is a block that performs more diverse tasks, such as error detection and correction, or smoothing of an obtained contour.

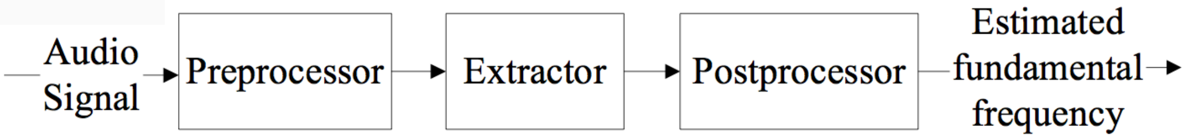

Figure 2.2 Steps of the fundamental frequency estimation process

2.2 Extractor methods

Current algorithms for estimating the fundamental frequency are presented in this section. Different classifications can be used for their categorization. The approach of distinguishing them by their processing domain is used here, separating them in time-domain and frequency-domain algorithms.

2.2.1. Time-domain algorithms

These algorithms try to find the periodicity of the input sound signal in the time domain.

a. Zero-crossing rate (ZCR)

ZCR is among the first and simplest techniques for estimating the frequency content of a signal in time domain, and consists in counting the number of times the signal crosses the 0-level reference (see Figure 2.3) in order to estimate the signal period. This method is very simple and inexpensive but not very accurate when dealing with noisy signals or harmonic signals where the partials are stronger than the fundamental. ZCR has also been found to correlate more with timbre than pitch and, thus, this method and its variants are not very much used for fundamental frequency estimation.

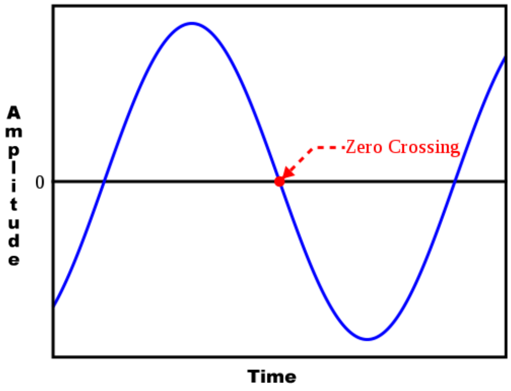

Figure 2.3 Zero-crossing of a signal

b. Auto-correlation function (ACF)

The ACF is a function that calculates the cross-correlation of a signal with itself, thus, returning the resemblance of a part of the signal with a preceding one. This allows finding the part that gets repeated, i.e. the period, that will correspond to the maximum value of auto-correlation (see Figure 2.4).

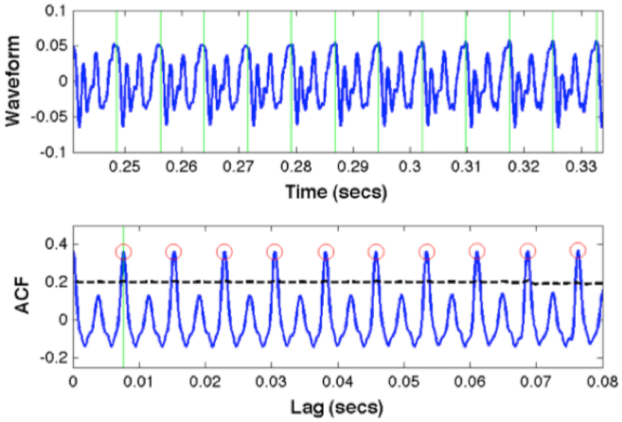

Figure 2.4 Waveform of a signal above, its corresponding ACF output below

ACF based algorithms [3], [4] have been among the most frequently used fundamental frequency estimators. For optimization reasons, they can also be computed in the frequency domain [5]. They have been reported to be relatively noise immune but sensitive to formants an spectral peculiarities of the analyzed sound [5]. Also, according to [5], ACF algorithms are more like to have “twice-too low” octave errors than “too high” octave errors, which are not probable.

In this kind of algorithms there must be conferred a special mention to the YIN algorithm [6] because of the wide use it has acquired nowadays. It is a modified version of the ACF that instead of using a multiplication for calculating the cross-correlation it uses the squared difference. This makes it robust to amplitude changes and solves the “too low” octave error. Moreover, it is a relatively simple method that may be implemented efficiently with low latency and has no upper limit for the f_0 search range.

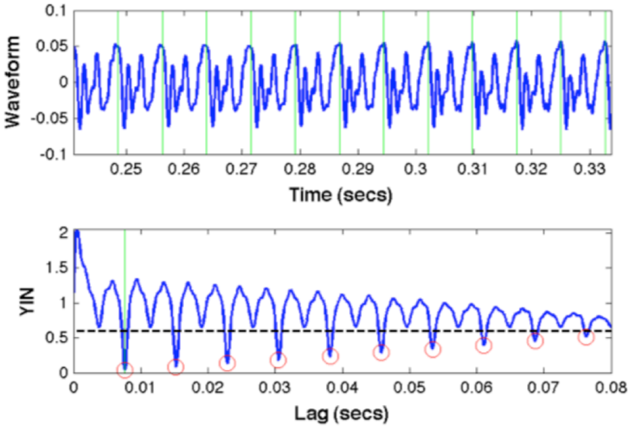

Figure 2.5 Waveform of a signal above, its corresponding YIN output below

c. Envelope periodicity

This model is based on the observation that signals with more than one frequency component exhibit periodic fluctuations (beatings) in its time domain amplitude envelope. The rate of these fluctuations depends on the frequency difference between each pair of frequency components. In the case of a harmonic sound, the f_0 interval outstands and the fundamental frequency is clearly visible in the amplitude envelope of the signal. Some estimators [7], [8] have included models of human pitch perception. These methods attempt to estimate the perceived pitch, not the pure physical periodicity.

d. Parallel processing

This model comes from [9] and [10], an algorithm purely based on time domain processing that has been used in a wide variety of applications. It is conceived in three steps (depicted in Figure 2.6):

1. The speech signal is processed to create a number of impulse trains that retain the periodicity of the original signal and discard features that are irrelevant to the pitch detection method. This can be considered as the pre-processing part of the algorithm.
2. Simple estimators are used to detect the period of the impulse trains.
3. All estimates are combined to infer the fundamental frequency of the speech waveform.

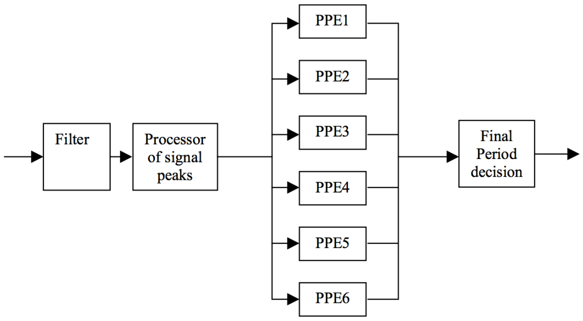

Figure 2.6 Steps of the parallel processing approach

This algorithm has the advantage that several different processes analyze in parallel the same problem, thus, when one fails the other ones still succeed. This redundancy strategy is believed to exist in the human perception [11]. This algorithm has a very low computational complexity and performs relatively well.

2.2.2. Frequency-domain algorithms

These algorithms estimate the fundamental frequency using the spectral information of the signal, obtained by the Discrete Fourier Transform (DFT) or another transform.

a. Cepstrum analysis

Cepstrum is the inverse Fourier transform of the logarithm of the power spectrum of the signal (see Figure 2.7). Thanks to the logarithm operation, the source and the transfer functions are separated. Consequently, the pulse sequence originating from the source periodicity re-appears in the cepstrum as a strong peak at “quefrency” (lag time) T_0. Cepstrum was introduced in [12] for determining the fundamental frequency of speech signals.

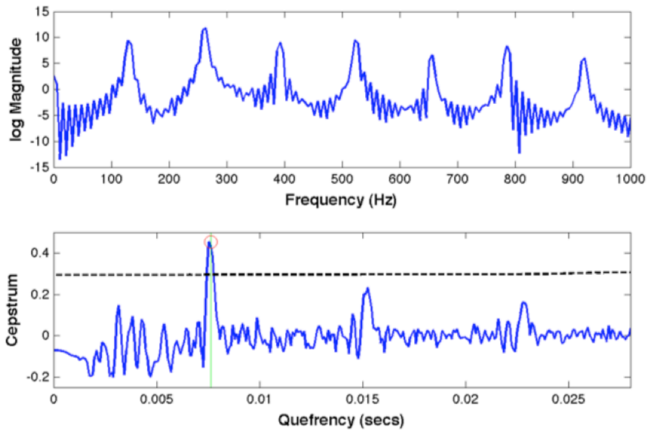

Figure 2.7 Logarithmic magnitude spectrum of a signal above, its corresponding cepstrum below

The cepstrum fundamental frequency estimation model is similar to autocorrelation systems, just that, while the frequency domain ACF is based on the logarithm of the magnitude spectrum, cepstrum uses the power magnitude. Cepstrum also shares the defect of committing “too low” octave errors. On the other hand, unlike ACF systems, cepstrum f_o estimators perform poorly in noise but have good performances with formants and spectral peculiarities [5].

b. Spectrum auto-correlation

These methods are based on the idea that a periodic but non-sinusoidal signal has a periodic magnitude spectrum, the period of which is the fundamental frequency. This period can be estimated with the ACF (see Figure 2.7). These algorithms are robust against “too low” octave errors since there is no spectral periodicity at half the fundamental frequency rate, but “too high” octave errors may occur [5]. An implementation of this method is described in [13].

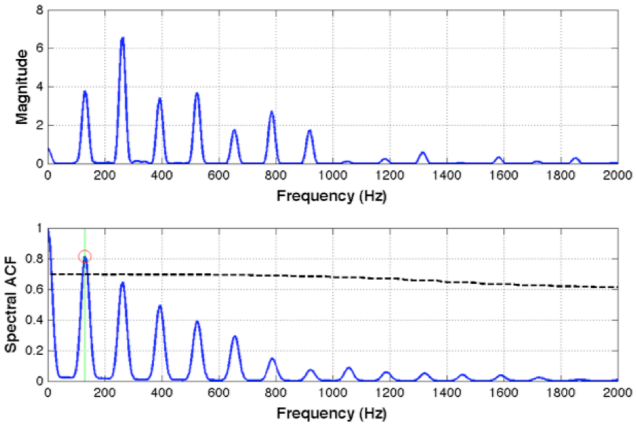

Figure 2.8 Magnitude spectrum of a signal above, its corresponding ACF below

c. Harmonic pattern matching

These algorithms are based on finding patterns on the magnitude spectrum of the signal. Their main idea comes from the fact that harmonic sounds present a regular structure, i.e. a pattern, which can be detected and used to estimate the fundamental frequency.

A first simple approach is to use comb filtering. A comb filter is band filter that repeats over all the frequency range, thus, it can be used to compute the energy for the different frequencies inside the f_0 search range in a way that harmonics will contribute when appropriate (see Figure 2.9). Comb filters are easy to implement and compute but any other filters with a different "template" can be used.

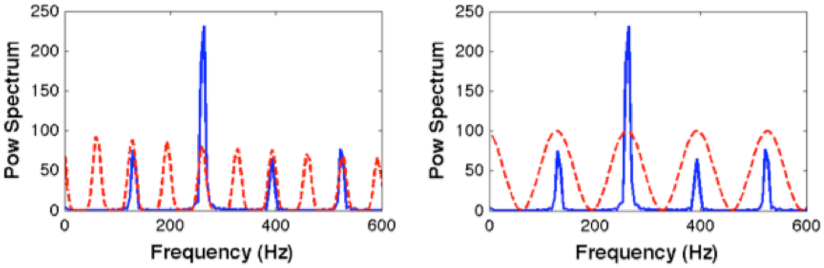

Figure 2.9 Harmonic pattern matching using comb filtering. In the left, a comb filter with half the f_0 is used. In the right, a comb filter with the correct f_0 is used.

A better approach for harmonic pattern matching consists on finding the spectral peaks in the magnitude spectrum. Then, a set of f_0 candidates is generated and the identified peaks are compared to the predicted harmonics (multiples of the f_0 candidate) for all candidates. The strategy used for comparing is called the *fitting measure*.

Some fitting measures were developed in [14] and [15], but the latter Two Way Mismatch method (TWM)[16] has acquired the most relevance. In this method, the discrepancy between the measured and the predicted sequence of harmonic partials is called the *mismatch error*. For each of the f_0 candidates, two mismatch errors are calculated; one from the measured partials to their nearest neighbor predicted partial, and the other from the predicted partials to their nearest neighbor measured partial (see Figure 2.10). This strategy avoids octave errors by applying a penalty for partials that are present in the measured data but are not predicted, and vice versa. It also benefits from the effect that any spurious components or partial missing from the measurement can be counteracted by the presence of uncorrupted partials in the same frame. The two mismatching errors as well as the final function for merging them are weighted with parameters empirically refined to make the procedure robust to the presence of noise or the absence of certain partials.

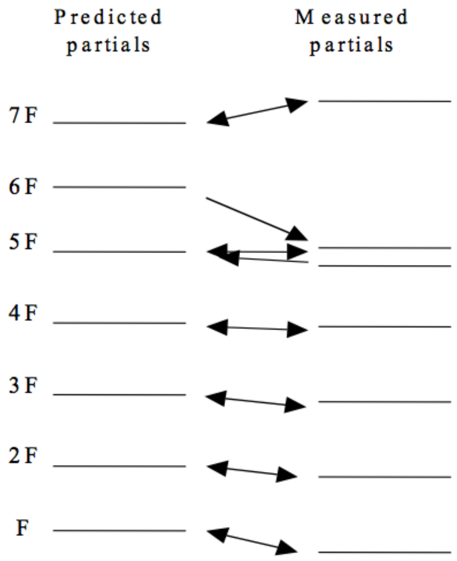

Figure 2.10 Two-way mismatch error method

d. Wavelet based algorithms

Some methods [17] try to take advantage of the wavelet transform (WT) to estimate the fundamental frequency. The main strength of the WT is that it performs a multi-resolution, multi-scale analysis that has been shown to be very well suited for music processing because of its similarity to how the human ear processes sound. In the STFT, which uses a single analysis window, a compromise between having enough resolution for low frequencies and not using a window to large to allow significant changes in high frequencies has to be done. In the other hand, WT uses short windows at high frequencies and long windows for low frequencies. That solves the problem having a good resolution for low frequencies while maintaining small enough windows for high frequencies.

e. Band-wise processing algorithms

Following the idea of constant frequency analysis using WT, [5] proposes an algorithm that calculates independent fundamental frequency estimates at separate frequency bands and combines them to yield a global estimate. This solves the problem of inharmonicity; partial intervals in inharmonic sounds are still constant at narrow enough bands. It also provides robustness against corrupted signals and interferences since these defects will be isolated at particular bands.

2.3 Voice/Unvoiced decision

Fundamental frequency estimation algorithms do not only have the mission of detecting the f_0 of a given signal, they also have to be able to distinguish segments that have f_0 (or pitch) versus those that do not, e.g. silences and percussion or noise segments, and consequently perform the f_0 estimation only for the pitched parts of the signal.

Typical techniques for accomplishing this distinction consist on using the estimated fundamental frequency itself, other measures derived from the method (e.g. the error measure in the TWM procedure) and relevant descriptors easy to compute (e.g. ZCR, energy, noisiness or harmonic distortion).

2.4 Multi-pitch estimation methods

Until now, only mono-pitch estimation methods have been presented. Those methods assume that a single fundamental frequency is present on the signal and, therefore, they return a single value for each time frame. Since there are situations were more than one pitch (being it from the same instrument or different ones) is present simultaneously, some methods for multi-pitch estimation have been proposed (a list of methods can be found in [18]).

This dissertation focuses on the study of the singing voice and as it fits the described case of mono-pitch, there is no need to review multi-pitch estimators. In the other hand, it is important to point out that because of the effects of reverb we will face situations were more than one fundamental frequency will co-exist at the same time (this will be explained in section 3.3.2.c). Even though this is a different scenario were we do not want to detect various fundamental frequencies but only the actual one, some ideas found in the literature of multi-pitch estimation could be useful to overcome this problem.

2.5 Estimators for the singing voice

Although the first algorithms to estimate the fundamental frequency for musical signals came from the speech literature [2], there are not specific monophonic estimators for the singing voice. Normally, the singing voice is considered as a special case of speech, where the same instrument is behaving a bit different. Because of that, the common speech estimators are used adapting them to the specificities of the singing voice (like in [19]), e.g. wider f_0 search ranges or specific pre-processing blocks for signal enhancement. Consequently, our study will consider speech specific algorithms for fundamental frequency estimation, like e.g. YIN [6].

In the other hand, we are actually aware of some f_0 estimation algorithms specifically designed to work with the singing voice but for contexts more similar to multi-pitch. One case is [20], that proposes an algorithm that performs singing voice f_0 estimation at the same time that separates this same singing voice from a music accompaniment (a task known as speech segregation). Another case is MELODIA [21], which is focused on extracting the main melody (the score, not the audio) of musical audio signals and, based on the assumption that the main melody in popular music is most of the times performed by a singer, it is optimized for the singing voice. Both methods use a process in which, following different strategies, a set of fundamental frequency candidates are extracted for every frame. Then, a decision function is responsible of choosing the more appropriate candidates, having in consideration the candidates of the neighboring frames. It is of our interest to also study the performance of these algorithms in our reverberation context and, after observing the results, analyze the possible advantageous strategies they have implemented.

2.6 Estimators for reverberated signals

In the later years there has been a lot of effort in improving and developing algorithms for reducing reverberation in signals, i.e. de-reverberation. Despite of that, few authors addressed the subject of robustly extracting the fundamental frequency of reverberated signals. This section will give an overview of the single proposed methods tackling with mono-pitch estimation that will be further studied and evaluated in the next sections of this dissertation.

Tomohiro Nakatani presented a f_0 estimator robust to background noise and spectral distortion [22] that was further exploited for his de-reverberation methods. It must be pointed out that he considers reverberation as one possible spectral distortion. Nakatani's method consists on a frequency method that creates a *dominance spectrum*, which is a spectrum where the amplitude is obtained from the instantaneous frequencies (IFs) [23] and a dominance measure that enhances the peaks of harmonics, suppresses variations produced by noise and whitens the spectral envelope eliminating spectral distortions (see Figure 2.11). Finally, a decision measure that summarizes the dominance of all harmonic components, called *harmonic dominance*, is used to determine f_0. The implementation was not publicly available and, thus, it could not be considered for our study.

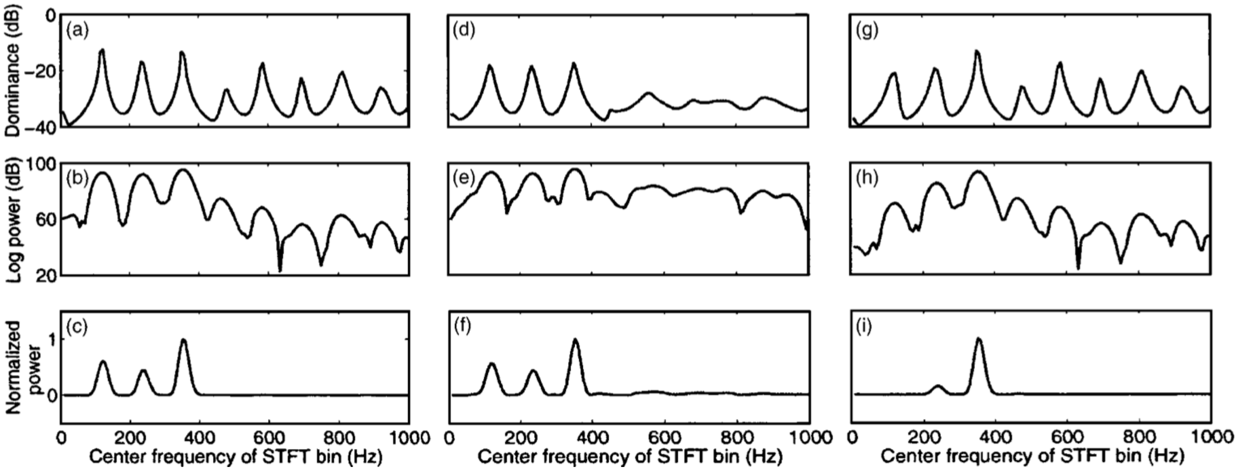

Figure 2.11 Dominance spectra [(a), (d), and (g)], logarithmic power spectra [(b), (e) and (h)], and power spectra [(c), (f), and (i)] of clean speech (left three panels), speech with white noise (middle three panels, SNR: 0 dB), and speech convolved with a SRAEN filter (right three panels)

3 Reverberation

Reverberation is the effect produced by the combination of acoustic reflections when sound waves propagate in enclosed spaces. Lets consider a single omnidirectional source of sound located within an enclosed space such as an office or living room with walls and other surfaces that reflect sound to some extent. If we assume that the source starts to emit at some instant in time $t = t_0$ and that the room was silent for $t < t_0$, then, the sound emanating from the source will be reflected multiple times in a manner that depends on the geometry of the source and the room as well as the nature of the reflective surfaces. This process produces a sound energy distribution that becomes increasingly uniform with time $t > t_0$ across a wide range of frequencies of interest.

This dissertation studies the effects of reverberation on fundamental frequency estimation algorithms. It is, then, capital to comprehend how reverberation works. This section presents an overview with the acoustic principles, models and measures for room reverberation. This builds the essential base of knowledge for the next subsection where current methods for speech de-reverberation, i.e. eliminating the effects of reverb from a speech signal, are presented. Finally, the effects that reverberation produces in a signal and the consequences they produce when estimating the fundamental frequency are described. This section is inspired by the introduction chapters in [24], which are recommended for further information.

3.1 Room acoustics

This section explains the basic physical processes of room acoustics, which are needed to know to understand reverb and de-reverberation methods.

3.1.1. Acoustical attributes

This section presents the attributes used for measuring reverberation.

a. Acoustic impulse response

The Acoustic Impulse Response (AIR) characterizes the acoustics of a given enclosure. Whereas AIR is used to refer to acoustic impulse responses in general, there are some cases where it is more appropriate to limit the acoustic context to be within a room, in which case, the impulse response is referred to as a Room Impulse Response (RIR). In this dissertation it will be used AIR and also RIR, depending on the acoustic scenario being considered.

The RIR is defined as the acoustic pressure pattern induced at a particular point in a room in response to a pressure impulse of unity magnitude at another point in the room. Between this to points, the room's acoustical properties can be seen as a linear and time-invariant filter that produces a known acoustical output for every input. In signal processing this filter is the impulse response, in this case, of the room. The impulse response of a linear system is the waveform that appears at the output of a system when a unit impulse (Dirac delta function) is presented at the input. The output $y(t)$ for arbitrary input $x(t)$ is obtained by convolving the input with the impulse response $h(t)$.

$$y(t) = x(t) * h(t) = \int_{-\infty}^{\infty} h(\tau) x(t-\tau) d\tau \tag{3.1}$$

The lower limit for integration is set to zero for physically realizable causal systems. The system is said to be BIBO stable (Bounded-Input, Bounded-Output) if the output is bounded for every bounded input. If the impulse response $h(t)$ does not change with time, the system is time-invariant. Finally, if the superposition principle holds, the system is linear. Systems that fulfill the two previous conditions are called Linear Time-Invariant (LTI) systems. When moving to the world of discrete-time systems, the convolution integral in (3.1) becomes a convolution sum,

$$y[t] = x[t] * h[t] = \sum_{k=-\infty}^{\infty} h[k] x[n-k] \tag{3.2}$$

where n is the discrete time index. The output sequence $y[n]$ is thus related to the input sequence $x[n]$ by a linear combination of the past and future values, the weights being given by the unit sample response $h[n]$. For causal systems the lower limit for the sum is zero.

Several models of room impulse response have been considered in the literature [25], [26], [27], [28], [29], [30], being the main ones Finite Impulse Response (FIR) systems and

Infinite Impulse Response (IIR) systems, where $h[t]$ does not depend on itself for the former but it does for the last. Typically, the AIRs are divided in two distinct regions; the early and the late reflections. The early reflections are often taken as the first 50ms of the impulse response [31], and constitute well defined impulses of large magnitude relative to the smaller magnitude and diffuse nature of the late reflections.

Figure 3.1 shows an example room impulse response. Direct-path propagation from the sound source to the microphone gives rise to an initial short period of near-zero amplitude, sometimes referred to as the direct-path propagation delay, followed by a peak. The amplitude of this peak due to direct-path propagation may be greater or less than the amplitude of the later reflections depending on the source-microphone distance and the reflectivity of the surfaces in the room. The example of Figure 3.1 shows a relatively strong direct-path component, indicating that the source- microphone distance is relatively short.

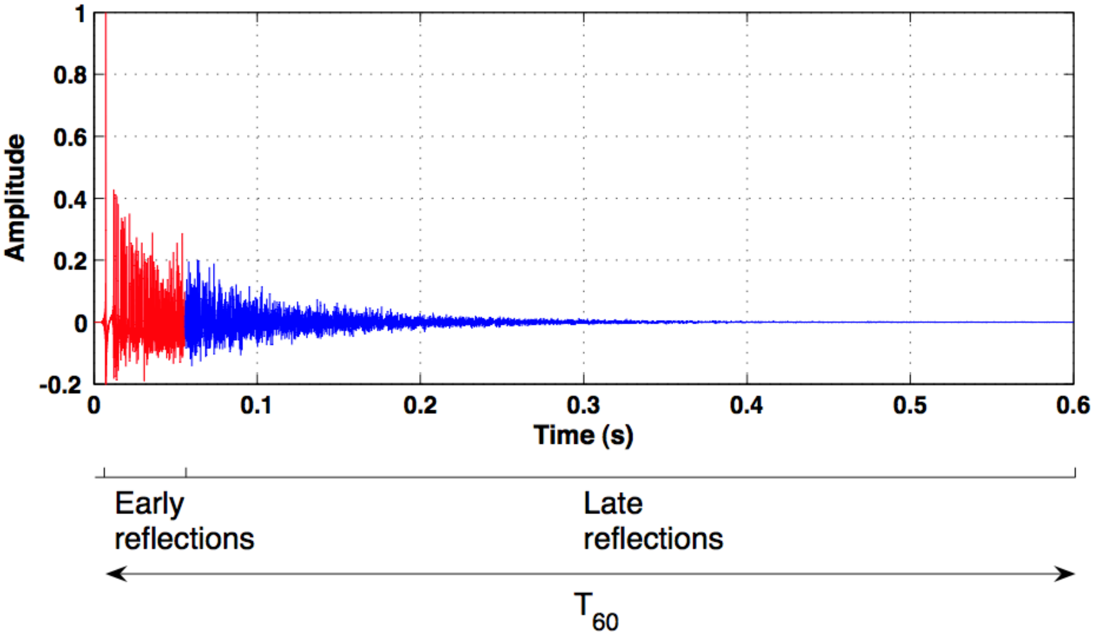

Figure 3.1 An example room impulse response

b. Wave-equation

In physics, a sound field can be understood as a superposition of plane waves. Typically, several simplifications are assumed: the medium in which the waves travel is homogeneous, it is at rest, and its characteristics are independent of the wave amplitude. Then, the propagation of such acoustic waves through a material can be considered to be a linear process and this propagation can be described by the (second order partial differential) wave equation. The wave equation describes the evolution of sound pressure $p(q,t)$, without any driving source, as a function of position $q=(q_x,q_y,q_z)$ and time t and is given by

$$\nabla^2 p(q,t) - \frac{1}{c^2}\frac{\partial^2 p(q,t)}{\partial t^2} = 0, \tag{3.3}$$

where

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{3.4}$$

The wave equation can be expressed in the frequency domain by taking the Fourier transform of sound pressure, $p(q,t)$, given by

$$P(q,\omega) = \int_{-\infty}^{\infty} p(q,t)e^{-j\omega t}\,dt, \tag{3.5}$$

to give the Helmholtz equation

$$\nabla^2 P(q,\omega) + k^2 P(q,\omega) = 0, \tag{3.6}$$

where

$$k = \frac{\omega}{c} = \frac{2\pi}{\lambda} \tag{3.7}$$

is the wavenumber, ω is the angular frequency and λ is the wavelength.

c. Sound field in a reverberant room

When sound is produced in a room or other reverberant environment, a listener will hear a mixture of direct sound and reverberant sound. The direct-path component is the sound that travels from the source to the listener without reflection whereas the reverberant component is the sound that travels from the source to the listener via one or more reflections. The effect of increasing the distance between the sound source and the listening location is to reduce the energy of the direct-path component. The energy of the reverberant sound is not in general affected by the source-listener distance but instead is dependent on the acoustic properties of the room.

The sound energy density, i.e. sound energy per unit volume, due to the direct-path component is given by

$$E_d = \frac{QW_s}{4\pi c D^2} \tag{3.8}$$

where W_s is the power output from the sound in watts, D is the distance from the source and Q describes the directivity of the source (e.g. for an omnidirectional source $Q=1$).

Similarly, the sound energy density due to the reverberant component is given by

$$E_r = \frac{4W_s}{cR}, \tag{3.9}$$

with the room constant, R, given by

$$R = \frac{\bar{\alpha}A}{1-\bar{\alpha}}, \tag{3.10}$$

where $\bar{\alpha}$ and A denote the average absorption coefficient of the surfaces in the room and the total absorption surface area, respectively.

It can be seen that the energy density of the reverberant sound is independent of the distance D, whilst the direct sound energy density is related to D by an inverse square law.

d. Reverberation time

The Reverberation Time (RT or T_{60}) is the time interval in which the reverberating sound energy, due to decaying reflections, reaches one millionth of its initial value, i.e. the time interval it takes for the reverberation level to drop by 60 dB [32].

In a diffuse sound field, the ideal room decay process exhibits a purely exponential decay curve

$$p^2(t) = p_0^2 e^{-kt} \tag{3.11}$$

where p_0 is the sound pressure at zero time, and $p(t)$ is the sound pressure at time t. The decay parameter k is related to the room properties by

$$k = \frac{cA}{4V} \tag{3.12}$$

where $c \simeq 340 m/s$ is the velocity of sound, A the total absorption area in the room, and V the volume of the room. According to Sabine [31], the reverberation time is defined as

$$T_{60} = \frac{0.16V}{A} \tag{3.13}$$

The coefficient 0.16 is empirically determined, and shows some variance with temperature. Combining equations (3.12) and (3.13), the parameter k is related to the reverberation time by

$$k = \frac{13.6}{T_{60}} \tag{3.14}$$

We can also define the reverberation time related with the average absorption constant, $\bar{\alpha}$, of the resonant modes in the room as:

$$T_{60} = \frac{3\ln(10)}{\bar{\alpha}} \tag{3.15}$$

Combining the equations (3.14) and (3.15) a relation between the average damping constant $\bar{\alpha}$ and k can be observed: $2\bar{\alpha} \approx k$. Furthermore, the reverberation time for a given room is seen from these expressions to be independent of the position within the room of the sound source and the measurement location.

e. Energy decay curve

Energy Decay Curve (EDC) is the decay of the squared sound pressure against time from the instant a broadband sound source is switched off after having obtained a steady state uniform sound energy distribution. If the Acoustic Impulse Response (AIR) of the room, $h(t)$, is known, the EDC can be obtained from the Schroeder integral [31]

$$EDC(t) = \int_t^{\infty} h^2(\tau)\delta\tau \tag{3.16}$$

In practice the upper limit of integration in (3.16) is set to a time instant at which the decay curve is still a little bit above the noise floor. The practical formula for obtaining the decay curve then becomes [33]

$$EDC(t) = N\int_t^{T_i} h^2(\tau)\delta\tau \tag{3.17}$$

where N is a constant proportional to the Power Spectral Density of the noise on the frequency range measured and T_i is the upper limit of integration. According to [33], the choice of T_i should be made so that it is close to the point where the decaying signal "dives" into the noise floor. ISO 3382 standard specifies that T_i should be set to a point where the impulse response is 5 dB above the noise floor.

Converting levels to a decibel scale, the decay can be described by a linear equation $y = ax + b$, where the decay curve of slope a is at level y at time x. Offset b is usually equal to zero, as the curve is commonly normalized to begin at a level of 0 dB, thus passing through the origin.

Figure 3.2 shows an example of the Schroeder integration curve calculated from a room impulse response. An example of straight-line fits to the first 30 *dB*.

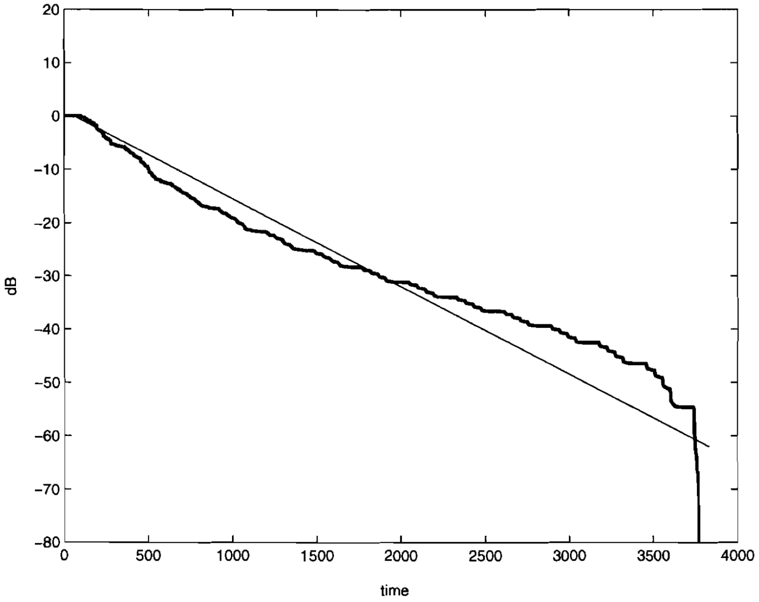

Figure 3.2 Energy Decay Curve (EDC)

f. Energy decay relief

To give additional insight, the impulse response can be split into frequency sub-bands and the EDC computed in each sub-band to give the Energy Decay Relief (EDR) as a function of both time and frequency, EDR(t, f) [31], [34]. This is typically presented as a 2-D surface plot and enables the frequency dependence of reverberation time to be studied.

g. Early decay time

Early Decay Time (EDT) is defined as the time interval required for the sound energy level to decay 10 dB after excitation has stopped. In a direct comparison with reverberation time the result is scaled by a factor of 6.

h. Critical distance

The critical distance is defined as the distance D_c from the source at which the sound energy density due to the direct-path component, E_d, and the sound energy density due to the reverberant component, E_r, are equal. It is evaluated by equating (3.8) and (3.9) to give

$$\frac{Q}{4\pi D_c^2} = \frac{4}{R}, \tag{3.18}$$

so that

$$D_c = \sqrt{\frac{QR}{16\pi}} \tag{3.19}$$

As shown in [31], the critical distance can also be expressed in terms of Q, V and the reverberation time T_{60} as

$$D_c \approx 0.1\sqrt{\frac{QV}{\pi T_{60}}} \tag{3.20}$$

An example of sound energy density in a room as a function of the distance from the source is shown in Figure 3.3.

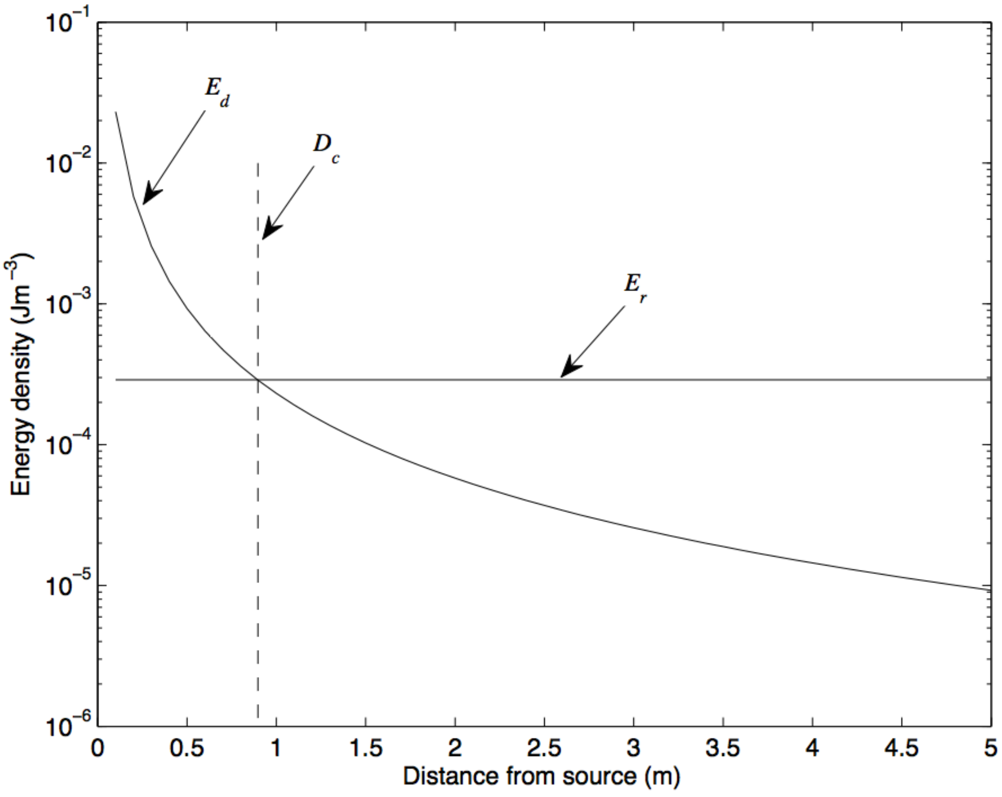

Figure 3.3 Direct energy density, E_d, and reverberant energy density, E_r, against distance from a 1 watt source in a room of dimensions 3×4×5m with $T_{60} \approx 0.29s$ (using the Eyring sound absorption coefficient $\alpha=0.3$) and $c=344m/s$. The vertical dashed line indicates the critical distance, $Dc \approx 0.9m$, computed using the approximate formula in (3.20).

3.1.2. Models of room reverberation

Three main models (see Figure 3.4) can be considered for modeling reverberation: first, wave based modeling which is based on the wave equation and is a fundamental approach, second, ray based modeling which leads to ray tracing methods and, third, the image method for modeling reverberation using virtual sources. Indeed, different analysis techniques are appropriate for different ranges of frequency of sound. Therefore, a combination of modeling techniques is necessary to achieve accuracy over the full audio spectrum. However, speech signals have a limited bandwidth that allows a simpler modeling.

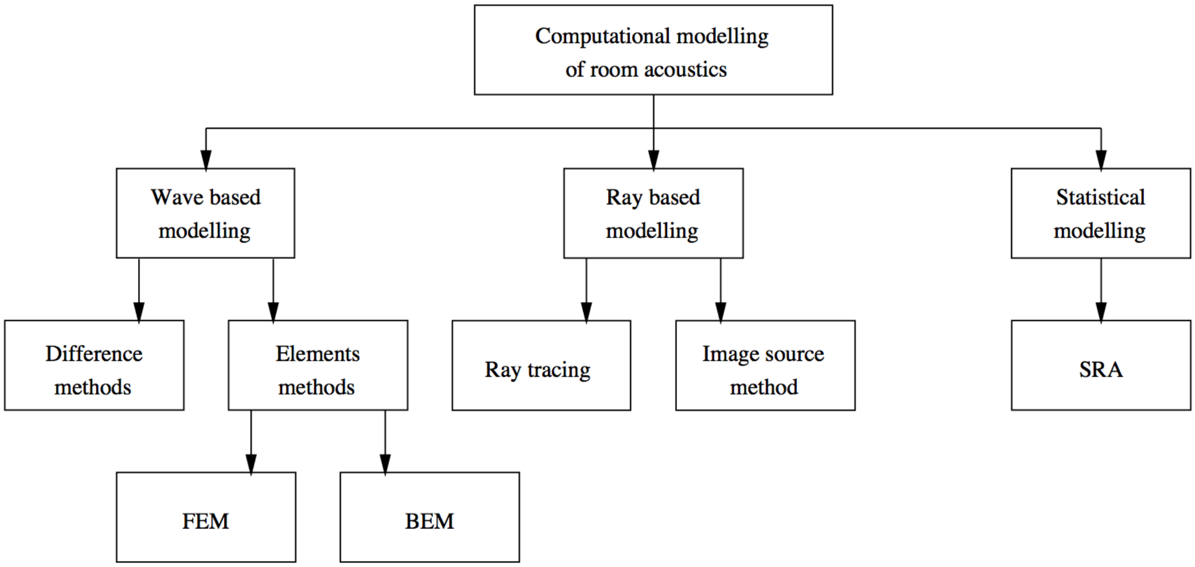

Figure 3.4 Methods for modeling and simulating room acoustics

3.2 Speech de-reverberation

De-reverberation is the process to reduce the reverberation effects in a signal, be it by means of processing the signal or re-synthesizing it. De-reverberation methods may be divided considering different classifications, for example, single vs. multi-microphone techniques, those primarily affecting coloration vs. those affecting late reverberation, or those that need to estimate the AIR vs. those that do not. This dissertation categorizes speech de-reverberation methods in two classes. The first approach, called *linear filtering*, de-reverberates time-domain signals or STFT coefficients. The second approach, called *spectrum enhancement*, de-reverberates corrupted power spectra while ignoring the signal phases.

It must be pointed out that even though the object of study of this dissertation is the singing voice, here we present de-reverberation speech methods. That is because currently no specific methods for the singing voice exist in the literature, thus, the singing voice will be considered as a particular case of speech.

Before surveying the three approaches in more detail, we summarize the notations. The STFT of reverberant signal $y(t)$ is denoted by $y_n[k]$, where n represents the index of a time frame and k refers to a frequency bin. The various representations of $x(t)$ are defined similarly.

3.2.1. Linear filtering

The linear filtering approach attempts to remove the effect of reverberation in the time or STFT domain taking consecutive reverberant observations into account. In contrast to the other approach, linear filtering exploits both the amplitudes and phases of the signal, which is advantageous in terms of accuracy because reverberation is a superposition of numerous time-shifted and attenuated versions of a clean signal so that both the amplitudes and phases are useful for de-reverberation. In addition, taking the signal phases into account enables us to effectively exploit the acoustical differences between multiple microphone positions [35], but this is not the case for this dissertation since it only focuses on single microphone methods as this is the typical set-up for capturing the singing voice and, thus, multi-microphone methods will not be considered; for more information see [24] and the references in there. Therefore, the algorithms of this approach are explained assuming a single microphone and the STFT representation, even though many of the algorithms can be extended to benefit from multiple microphones as discussed in [24].

To represent the relationship between clean and reverberant STFT coefficients, $x_n[k]$ and $y_n[k]$, the following representation is often assumed in the literature [36], [37]:

$$y_n[k] \approx \sum_{\tau=0}^{T} h_\tau[k]^* x_{n-\tau}[k], \tag{3.21}$$

where the superscript * stands for complex conjugation and T is the number of time frames over which reverberation continues to have an effect. The complex conjugate of $h_n[k]$ is used for consistency with the notation commonly accepted in the field of adaptive filtering [38]. Equation (3.21) means that the effect of reverberation may be represented as a one-dimensional convolution in each frequency bin, and therefore sequence $(h_n[k])_{0 \le n \le T}$ can be viewed as an STFT-domain counterpart of the time-domain room impulse response. The objective is to recover the corresponding clean STFT coefficients $(x_n[k])_{n \in \mathrm{T}}$ for each k, given a sequence of reverberant STFT coefficients $(y_n[k])_{n \in \mathrm{T}}$. Below, the frequency bin index k is omitted for conciseness.

As the name suggests, linear filtering methods employ a linear filter to perform dereverberation according to

$$x_n = \sum_{\tau=\mathrm{T}_\perp}^{T_\perp} g_\tau^* y_{n-\tau}, \tag{3.22}$$

where $G = \{g_\tau\}_{T^- \le \tau \le \mathrm{T}^+}$ is a set of adjustable linear filter coefficients. Generally, $T^- \le 0$ and $T^+ > 0$. The clean STFT coefficient x_n is estimated based on $T^- + T^+ + 1$ consecutive reverberant frames, and thus the linear filtering methods naturally allow us to take the long-term acoustic context into account. The goal is to find an optimal filter G that cancels the room impulse h_n. Denoting the convolution of the room impulse response and the linear filter $f_n = \sum_{\tau=\mathrm{T}^-}^{T^+} g_\tau^* h_{n-\tau}$, the objective is to set G so that f_n is nonzero if $n = 0$ and zero otherwise. This problem is called blind de-convolution and has been studied extensively, especially in the field of digital communications [38]. Different blind de-convolution methods for speech signals are discussed in [27], [39].

In the following, we look more closely at blind de-convolution based on long-term linear prediction [36], [40], [41], [42]. It leverages an explicit speech model to determine the filter G. In one exemplary concept [36], which has been applied to various speech recognition tasks including meeting recognition, the speech model defines the probability density function (pdf) of a clean STFT coefficient x_n and is assumed to be a normal distribution with zero mean and variance θ_n. The time varying modeling, i.e., the dependence on frame index n, of the variance was shown to play a critical role in precise adjustment of the filter coefficients [36]. Since $\Theta = \{\theta_n\}_{n \in \mathrm{T}}$ is unknown in advance, the filter G is optimized jointly with Θ by using the method of maximum likelihood. Specifically, the likelihood of the combination of G and Θ given the sequence $Y = (y_n)_{n \in \mathrm{T}}$ of observed reverberant STFT coefficients is maximized according to

$$(\hat{G}, \hat{\Theta}) = \underset{(G,\Theta)}{\arg\max} \log p(Y \mid G, \Theta). \tag{3.23}$$

To facilitate the definition of the pdf $p(Y \mid G, \Theta)$, the concept of multistep prediction [42] is introduced. With multistep prediction, it is assumed in (3.22) that $g_0 = 1$ and that $g_n = 0$ when $T^- \leq n < T_\delta$, where T_δ is a positive integer that approximately corresponds to the boundary Δ between early reflections and late reverberation. This approach is called multistep prediction because, with these assumptions, (3.22) can be rewritten in the form of long-term T_δ-step forward prediction of y_n as

$$y_n = x_n + \sum_{\tau=\mathrm{T}_\delta}^{T_\delta} g_\tau^{*} y_{n-\tau}, \tag{3.24}$$

representing the current reverberant observation y_n as the sum of the clean signal x_n and a signal predicted from past observations with filter $G = \{g_n\}_{T_\delta \leq n \leq T^+}$. The sign of g_t has been inverted when deriving (3.24) from (3.22). Thanks to the predictive form of (3.24), $p(Y \mid G, \Theta)$ can be easily defined and the optimization problem in (3.23) is finally rewritten as the following minimization problem:

$$(\hat{G}, \hat{\Theta}) = \underset{(G,\Theta)}{\arg\min} \sum_{n \in \mathrm{T}} \left(\frac{\left| y_n - \sum_{\tau=T_\delta}^{T^+} g_\tau^{*} y_{n-\tau} \right|^2}{\theta_n} + \log \theta_n \right), \tag{3.25}$$

which can be solved by an iterative algorithm updating estimates of G and Θ alternately [36]. If multiple microphones are available, (3.24) is modified so that the current reverberant observation at a microphone is predicted from past observations from all the microphones, i.e., (3.24) is rewritten in the form of multi-channel prediction [36], [42].

The long-term prediction method has been successfully applied to actual meeting data. Furthermore, by extending the observation pdf $p(Y \mid G, \Theta)$, this method can be modified to deal jointly with multiple speakers, additive background noise and reverberation as described, for example, in [43].

3.2.2. Spectrum enhancement

As an alternative to linear filtering, enhancement may be performed after taking the squared magnitudes of the STFT coefficients. The objective of the resulting spectrum enhancement methods is to restore the clean power spectrum coefficients $(|x_n[k]|^2)_{n\in\mathrm{T}}$, given a sequence of the corresponding reverberant power spectrum coefficients $(|y_n[k]|^2)_{n\in\mathrm{T}}$. The advantage of spectrum enhancement over linear filtering is its high robustness against speaker movement, which derives from the fact that the magnitude of the late reverberation is largely insensitive to changes in speaker and microphone positions. Furthermore, spectrum enhancement methods can be easily combined with conventional additive noise reduction techniques, such as spectral subtraction, as shown in [44].

The spectrum enhancement methods can be categorized into two classes according to the estimator of the reverberation power spectrum: moving-average estimator and predictive estimator. The moving-average estimator is based on the power spectrum-domain reverberation model given by

$$|y_n[k]|^2 \approx \sum_{\tau=0}^{T} |h_\tau[k]|^2 |x_{n-\tau}[k]|^2 , \tag{3.26}$$

which is derived from (3.21) by disregarding the cross-terms between different time frames. To estimate the power spectrum of late reverberation or clean speech with this model, we need to know the power spectrum-domain representation $(|h_n[k]|^2)_{0\le n\le\mathrm{T}}$ of the room impulse response. This can be achieved by techniques such as correlation analysis [45], non-negative matrix factorization [46], [47], and an iterative least squares method [48].

The predictive reverberation estimator employs a much simpler model [44], [49]. Assuming a strict exponential decay of the late reverberation magnitude, the power spectrum $|r_n[k]|^2$ of the late reverberation at frame n can be predicted from the power spectrum $|y_{n-T_\delta}[k]|^2$ of the reverberant observation at frame $n-T_\delta$ via a scalar predictor $a[k]$ as

$$|r_n[k]|^2 = a[k] |y_{n-T\delta}[k]|^2 \tag{3.27}$$

T_δ is set at a value corresponding to approximately 50 ms. The predicted late reverberation is removed from the reverberant power spectrum $|y_n[k]|^2$ with spectral subtraction. The predictor $a[k]$ is determined based on the knowledge of T_{60}.

3.3 Effects of reverberation on voice analysis

This section describes the effects that reverberation produces on a clean anechoic speech signal. First, the general effects are presented, those perceivable from the signal point of view, and then those indirectly produced when we try to estimate the fundamental frequency.

3.3.1. General effects

Reverberation is a superposition of numerous time-shifted and attenuated versions of a clean signal. The interaction of this versions, or repetitions, produces mainly two effects:

a. Spectral coloration

The spectral coloration is the effect of changing the color (quality, timbre) of the original signal. Using the Fourier transform in equation (3.2) to move to the spectral-domain

$$Y[k] = X[k]H[k] \tag{3.28}$$

it can be seen that the impulse response $h[n]$ is equivalent to the filter $H[k]$ and thus, reverberation produces the same effect an LTI system could produce on the signal; it causes spectral changes and lead to a perceptual effect referred as coloration [31]. This effect is typically associated to the early-reflections part of the impulse response (around the first 50 ms)

b. Temporal tail

The late reflections of the impulse response (typically 50 ms after the direct-path signal) are referred to as the tail of the impulse response and constitute closely spaced, decaying pulses, which are seemingly randomly distributed. The late reflections cause a ‘distant’ and ‘echo-ey’ sound quality we refer to as the reverberation tail and provide the major contribution to what is generally perceived of as reverberation in everyday experience.

A graphical example can be seen in the spectrogram of a sweep (a pure sine wave with linearly increasing frequency over time) in Figure 3.5. When the sweep is reverberated its lower frequency in past samples keeps sounding for a while with decreasing intensity.

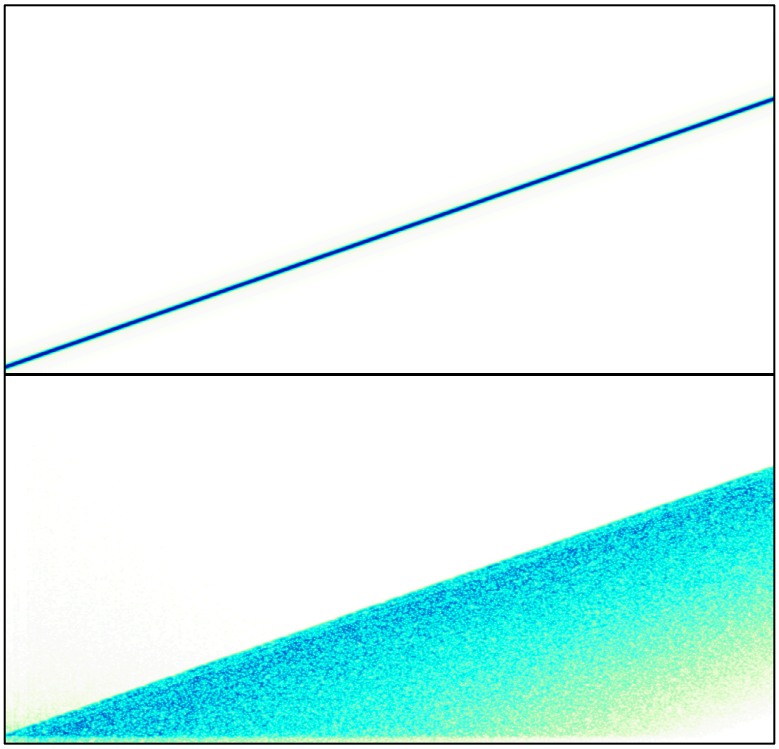

Figure 3.5 Sweep spectrogram. Clean signal above, reverberated signal below. Horizontal axis represents time and vertical axis frequency.

3.3.2. Fundamental frequency estimation effects

The general effects presented on the previous section have negative consequences when trying to estimate the fundamental frequency of speech or the singing voice. There is little literature [50] that studies the effects of reverberation on fundamental frequency estimators because most of the research has been focused only on automatic speech recognition (ASR), but the results found are highly coincident: in summary, detrimental effects are observed after a certain early reflections time limit around 25-50ms. This means early reflections are less problematic because they usefully increase the level of the speech and introduce spectral distortions that can be addressed using within-frame processing schemes (such as cepstral mean normalization). On the other hand, late reflections have a temporal smoothing effect that extends over several frames presenting a particular problem for F0 estimators. In other words, spectral coloration, despite changing the harmonic amplitudes of the signal, does not produce any relevant change for most estimators. And the temporal tail causes important disorders to estimation algorithms, which will decrease its accuracy as the late reflections acquire more length.

a. Unvoiced frames become voiced

The unvoiced frames are those where there is speech but it has no fundamental frequency, such as when there are unvoiced consonants. Because of the effects of the reverberation tail, vowels endure longer in time. This produces that when an unvoiced consonant is after a vowel, or even a voiced consonant, it can be 'shaded' by them and their frequency would be detected. Figure 3.6 illustrates an example with the waveform in red, the spectrogram with the white-yellow-blue range of colors, on top of it a dark blue line for the detect fundamental frequency and with red circles the zones of interest where the described effect happens. A clean singing voice signal has been used and it is shown in the upper part. After applying a reverb, depicted in the part below, we can observe how the algorithm is assigning a frequency to frames where there was not before and that corresponded to unvoiced consonants.

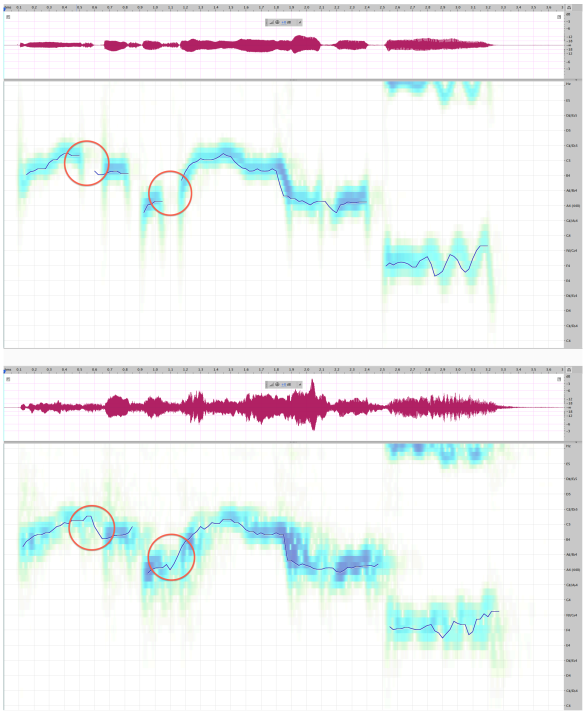

Figure 3.6 Effects of reverb in fundamental frequency estimation. Unvoiced frames become voiced

b. Silence becomes voiced

Similarly to the preceding effect, the voiced frames of vowels or voiced consonants endure in time, and this produces that, if there is a silence at the end of words or sentences, it will be detected with the frequency of the preceding frames. The same example as in Figure 3.6 is now presented in Figure 3.7 but with a single circle pointing to the end of a word where a silence has 'acquired' a the preceding frequency.

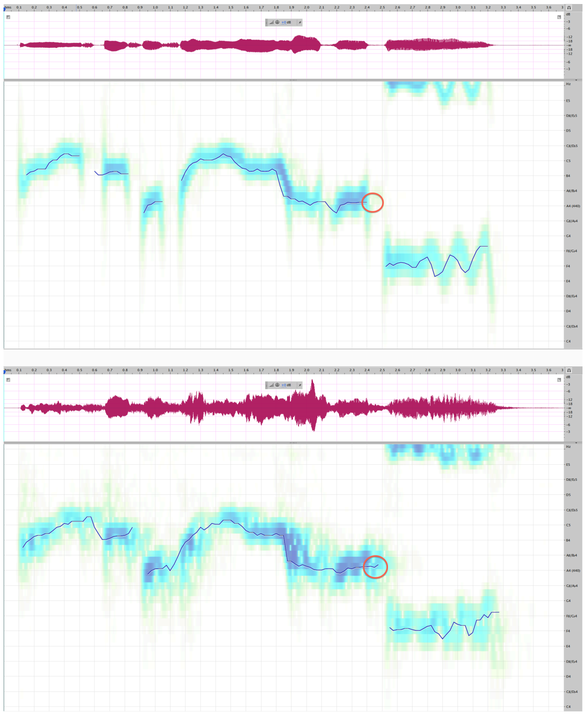

Figure 3.7 Effects of reverb in fundamental frequency estimation. Silence becomes voiced

c. Monophonic becomes polyphonic

This is probably the most challenging effect to overcome for estimators but, luckily, greater reverberation times are normally required, compared to the already discussed effects, for the problem to appear. Basically, the fact that a certain frequency persists more in time than its original clean version (see Figure 3.5) causes that, in a particular instant, more than one fundamental frequency will co-exist. For example, if the current pronounced vowel is a C3 note, we could also be hearing the preceding vocal, for instance A3. If it happens that they have different fundamental frequencies a previous monophonic situation with a single source has become equivalent to a polyphonic situation where more than one fundamental frequency is present at the same time. Because the estimator is expecting a singing voice, thus, a monophonic source, it will not be prepared to deal with simultaneous fundamental frequencies and, consequently, its estimation will start failing.

It is certain that, often, the preceding note, A3 in our example, will be in the background, i.e. C3 will be more loud and it will sound on top of A3. This is why estimators that consider predominance in amplitude, are less affected by this issue.

The same example as in Figure 3.6 and Figure 3.7 is presented in Figure 3.8. This time a reverberation with longer decay has been applied. It can be observed inside the red circle how a previous decaying melody becomes the first note enlarged.

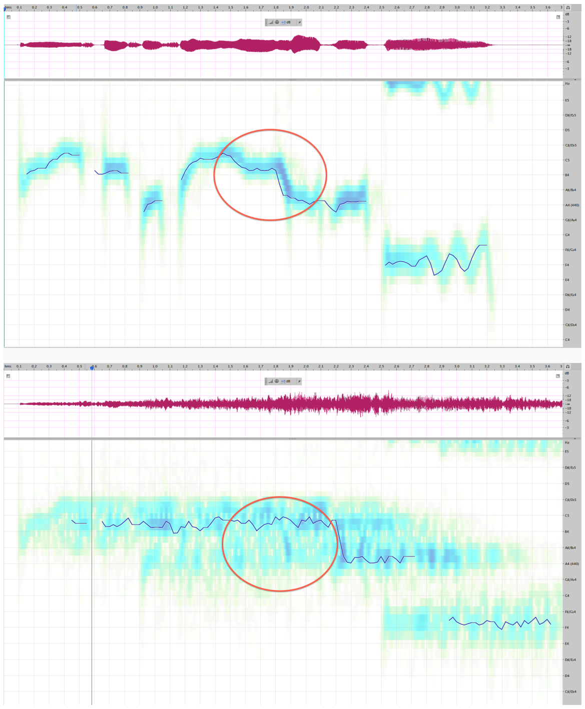

Figure 3.8 Effects of reverb in fundamental frequency estimation. Monophonic becomes polyphonic

d. Expressive features are smeared

The effects explained above produce the smearing of characteristic expressive resources used in the singing voice. Smearing is referred as loosing precision and, consequently, obtaining results less accurate where this expressive features are incorrectly detected, less notorious or not present at all. Typical effects that will be smeared by reverberation are:

1. **Vibrato**: it consists on modulating the frequency of a note. When reverb is applied really close fundamental frequencies will be co-existing during a note with vibrato. From the spectrum point of view, this will produce that really close main lobes (for the different fundamental frequencies) will be summed up indistinguishably forming a

wider main lobe, probably leading to less precision when picking up the its peak. Thus, even if a predominant amplitude is supposed for the current fundamental frequency, the modulation of the vibrato will be detected smaller than the real one. It could also happen that those really close fundamental frequencies are seen as a single fundamental frequency by the estimator, something similar to a unique fundamental frequency result of the mean of all of them, and no vibrato is detected at all but a plain note.

2. **Portamento**: it consists on the pitch sliding from one note to another (also referred as glissando). This feature suffers the same problem as vibrato; really close fundamental frequencies will be co-existing. A continuous case of portamento is a sweep (see Figure 3.5).
3. **Attack**: it consists on how the start of a note, in terms of amplitude, is produced. When reverberation is present changes in amplitude are generally smeared, especially if the critical distance is surpassed. This can lead to detecting the start of the note imprecisely. Moreover, when introducing a reverb on a clean signal, the direct-path introduces a constant delay. This fact will have to be considered when evaluating with a ground truth.

4 Study

In this section, the approach to solve the questions posed (see Section 1.2) in this thesis is presented.

4.1 Overview

A schematic view of the approach used in this study is presented in Figure 4.1.

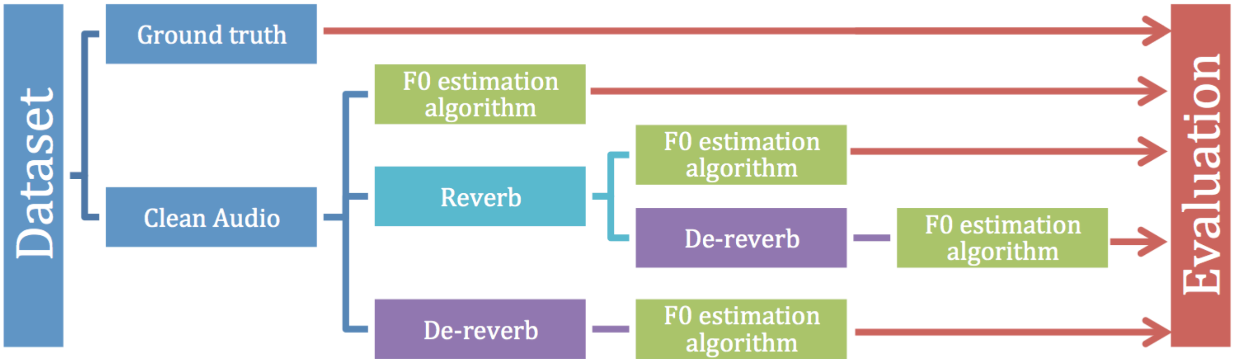

Figure 4.1 Schematic overview of the study

First, it is needed to choose a dataset to perform the study on. This dataset must provide, in one hand, clean excerpts of singing voice audio and, on the other hand, a ground truth indicating the fundamental frequency corresponding to the different frames of the excerpts. It is of utter importance that the audio is clean in terms of being dry, i.e. in reverb-free conditions. Second, it will be selected a set of fundamental frequency estimation algorithms that are considered of interest for our purpose and an evaluation method to contrast their output with the ground truth. Finally, every F0 estimation algorithm will be run and evaluated using: first, clean audio, second, clean audio reverberated, third, reverberated audio being de-reverberated and, eventually, clean audio being de-reverberated.

This methodology allows extracting different information in every stage. First, when considering only clean audio, the most accurate algorithm in normal conditions is found but also it allows us to decide if the chosen dataset is good enough (e.g. if it is clean enough) depending on the final accuracies obtained. Second, when evaluating the reverberated audio, the most robust algorithm against reverberation is discovered by means of observing the accuracy drop compared to clean audio results. Third, when evaluating reverberated audio that is de-reverberated, it is observed if some de-reverberation method can be used to improve the accuracy of the algorithms. Finally, the stage that runs clean audio directly through de-reverberation methods unveils if these de-reverberation methods have any deleterious effects in the F0 estimation algorithms accuracy in reverb-free conditions.

4.2 Audio dataset: MIR-1K

In the previous section it was explained we first needed a dataset with singing voice excerpts that were dry enough for our task. The main concern is that they are noise-free, so there is no interference for the F0 estimators, and they are reverberation-free, so we can artificially add it later to observe its effects. At present days there is a significant amount of speech datasets but not much with singing voice and most of them are small, lack F0 ground truth information or are not much clean. Luckily, a complete dataset with all these requirements was available and accessible for free: MIR-1K [51]. Here not many arguments can be added to the decision of choosing MIR-1K over other datasets since it could not be found any other one that satisfied totally all our requirements, but MIR-1K fit perfectly for our purposes.

MIR-1K is a dataset created by Chao-Ling Hsu and Jyh-Shing Roger Jang conceived for singing voice separation from music [52]. Even though it is not our task, its excerpts have the singing voice and music separated in channels. Actually, its excerpts were composed using karaoke Chinese pop songs as the background music and then they recorded its lab-mates singing over those songs. It contains 1,000 excerpts that range from 4 to 13 seconds and provide a total length of the dataset of 133 minutes. Having such a big dataset provides a high confidence in the results extracted from it. The singers are 8 males and 11 females, a fact that provides a good variety in pitch range, but most of them are amateur. This is acceptable since pitch contours of amateur singers will be more unstable and complicated to track for F0 estimators, thus the idea is that if it works with amateurs it is highly probable that will work with professionals and not the other way around, but a comparison with professional singers would be desirable for some applications. MIR-1K also provides a manually labeled ground-truth with pitch contours and voiced/unvoiced frames information. The audio quality, 16kHz and 16bit, is good enough since it covers the human singing pitch range. Besides of that, we also studied the noise conditions of the recordings. Recording quality does not comply with studio standards for many of the excerpts. Most of them have noticeable ground noise, they may have clicks and clip at some points. These interferences are not strong enough in the recordings to disturb a human in the task of extracting the pitch but it is known they could affect computer algorithms to do so. It will be shown later (in Section 5.1) that an accuracy of more than 90% is achieved with current F0 estimators in MIR-1K dataset, thus the interferences are not significant to alter our results significantly. This is probably due to the big amount of data in MIR-1K and the robustness of current methods to some of this interferences (i.e. background noise).

It is also interesting to point out that MIR-1K is one of the datasets used in the annual MIREX contest since 2009. This gives it an important credibility since many researchers in the field of MIR have used it. In 2009 it was created a special version of MIR-1K to comply with their requirements, e.g. 10ms frame for pitch information. Since this pitch information was calculated using linear interpolation from the original 20ms frame pitch information [53], this study uses the original dataset.

As pointed before, MIR-1K has a huge amount of excerpts (1,000). This is generally an advantage because it gives reliability to the results but highly increases the computation time. In this study a smaller dataset had to be created out of MIR-1K that was called MIR-1K-Micro; there were algorithms whose computation time was practically endless and others that had to be run manually file-by-file. MIR-1K-Micro was created randomly selecting a single excerpt for every singer, thus obtaining a dataset with 19 excerpts. It was checked that this smaller dataset is still representative of the original one; it gave very similar results compared to the complete MIR-1K. In this study the complete MIR-1K results are provided by default and MIR-1K-Micro results appear only were the full dataset could not be used.

Another interesting fact to consider when using datasets for F0 extraction is its voiced/unvoiced frames percentage that is shown in the Figure 4.2 below. The first thing to notice is how the MIR-1K-Micro dataset preserves the same relation between percentages; a fact that allows us to directly compare MIR-1K results with MIR-1K-Micro results with confidence. The other issue is that there is a significant percentage of voiced frames compared to unvoiced ones. This means that, for example, a no-analysis strategy for an algorithm to discriminate voiced/unvoiced frames could be to label everything as voiced and it would obtain a 70% of accuracy. Since a voicing accuracy of 90% or more is desirable to obtain good results it can be said that the voiced/unvoiced frame factor in MIR-1K is valid for our study. Nevertheless, the ideal number of this factor should be similar to the data analyzed in a final application and this will always depend on the final application. This study is focused towards a general idea but not a final application, thus there is not an exact number to pursue.

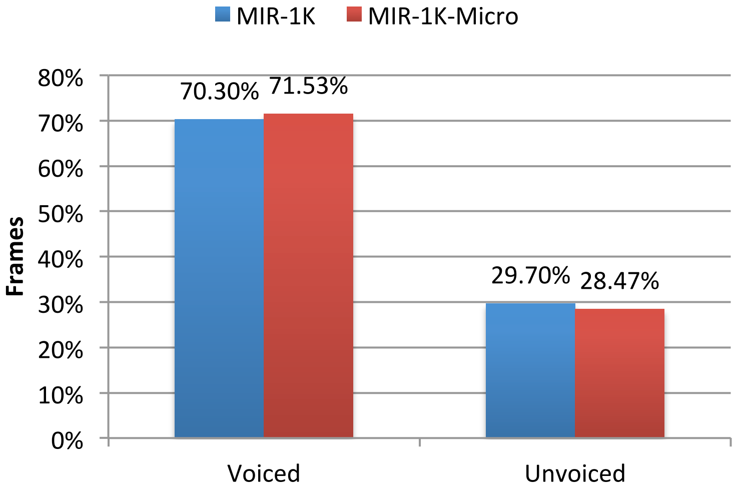

Figure 4.2 Percentage of voiced/unvoiced frames

4.3 Evaluation: MIREX audio melody extraction

To evaluate the results of the F0 estimation algorithms it is necessary some way to compare them with the ground truth and obtain a score that can be finally used to compare scores between different algorithms. For this study the audio melody extraction evaluation method from MIREX [54] was chosen. It was considered that the most standard method in the field was preferable over creating a specific one since it facilitates future comparisons with other studies. MIREX is an annual contest where different algorithms compete to get the best results in different categories of MIR research of current interest. The aim of the MIREX audio melody extraction evaluation is to identify the melody pitch contour from polyphonic musical audio. Pitch is expressed as the fundamental frequency of the main melodic voice, and is reported in a frame-based manner on an evenly-spaced time-grid. Although in this study there is no background music the evaluation method of MIREX audio melody extraction is perfectly suitable for our task. Indeed, this category evaluates monophonic F0 estimators just with the addition of background interferences (music in this case) because without them the problem is considered "solved". Moreover, reverberation can be seen as an interference of a different kind and makes sense to evaluate it with the same methodology.

4.3.1. Evaluation procedures

MIREX audio melody extraction evaluation consists of two parts:

- Voicing detection: deciding whether a particular time frame contains a melody pitch or not.
- Pitch detection: deciding the most likely melody pitch for each time frame.

There is a global score but these two parts are evaluated independently. This allows analyzing results with more detail and discriminating which of the parts is failing.

For the evaluation of the voicing detection of frames the following matrix is considered,

		Detected		
		unvoiced	voiced	sum
Ground truth	unvoiced	TN	FP	GU
	voiced	FN	TP	GV
	sum	DU	DV	TO

Table 4.1 MIREX evaluation procedure matrix

where the acronyms mean:

- TP (True Positives): voiced frames correctly detected.
- TN (True Negatives): unvoiced frames correctly detected.
- FP (False Positives): unvoiced frames incorrectly detected as voiced.
- FN (False Negatives): voiced frames incorrectly detected as unvoiced.
- GU (Ground truth Unvoiced) = TN + FP
- GV (Ground truth Voiced) = FN + TP
- DU (Detected Unvoiced) = TN + FN
- DV (Detected Voiced) = FP + TP
- TO (Total number of frames) = DU + DV = GU + GV

For the evaluation of the pitch detection the fundamental frequency detected has to be less than a half semitone close to the F0 ground truth. It does not matter how the algorithm labeled the frame in terms of voicing to check the pitch detection. The same rules apply for Chroma detection, only that all notes are fit in a single octave to no consider octave errors. Thus the there are the following breakdowns:

- TP = TPC + TPI
- TP = TPCch + TPIch
- FN = FNC + FNI
- FN = FNCch + FNIch

where the acronyms mean:

- TPC: True Positives Correct (they have a correct pitch)
- TPI: True Positives Incorrect
- TPCch: True Positives Correct chroma
- TPIch: True Positives Incorrect chroma
- FNC: False Negatives Correct
- FNI: False Negatives Incorrect
- FNCch: False Negatives Correct chroma
- FNIch: False Negatives Incorrect chroma

All the evaluation procedures presented above are then used to obtain the final evaluators. The evaluators currently used in MIREX [54] have been included (with the exception of voicing d-prime because it is not specified how to obtain it) as well as other evaluators used in older editions [55] for the sake of completeness of the study. Even though most of the evaluators are not needed for the analysis required in this study, all results are provided in the results appendix for future use in other research.

It is important to point out that when averaging pitch statistics in MIREX the performance of each excerpt is calculated individually and then the average of these measures is reported while, for the voicing detection, the average is simply made over all frames directly. They argue that, in the case of pitch, it helps increasing the effective weight of some minority genres that have shorter excerpts and, in the case of the voicing, some excerpts have no unvoiced frames and that can give misleading results. In the results analysis of this study it has been only used the averaging per excerpt. Indeed, MIREX audio melody extraction is using more datasets than only MIR-1K, which does not have these problems. Besides, we had calculated results with both strategies and there was no significant difference.

4.3.2. Voicing evaluation

Voicing evaluators used are the following:

- **Voicing Detection**: probability that a frame which is truly voiced is labeled as voiced (also known as "hit rate")

$$\frac{TP}{GV} \tag{3.29}$$

- **Voicing False Alarm**: probability that a frame which is actually unvoiced is labeled as voiced

$$\frac{FP}{GU} \tag{3.30}$$

- **Voicing Accuracy**: probability that the voicing of a frame is labeled right

$$\frac{TP+TN}{TO} \tag{3.31}$$

4.3.3. Pitch evaluation

Pitch evaluators used are the following:

- **Raw Pitch Accuracy**: probability of a correct pitch value (to within ±¼ tone) given that the frame is indeed pitched. This includes the pitch guesses for frames that were judged unvoiced

$$\frac{TPC+FNC}{GV} \tag{3.32}$$

- **Raw Chroma Accuracy**: probability that the Chroma (i.e. the note name) is correct over the voiced frames. This ignores errors where the pitch is wrong by an exact multiple of an octave (octave errors)

$$\frac{TPCch+FNCch}{GV} \tag{3.33}$$

4.3.4. Overall evaluation

Evaluators used that consider both voicing and pitch are the following:

- **Overall Accuracy**: probability that a frame is correctly labeled with both pitch and frame
$$\frac{TPC+TN}{TO} \tag{3.34}$$
- **Voiced Pitch Accuracy**: probability that a truly voiced frame is correctly labeled as voiced and the right pitch
$$\frac{TPC}{GV} \tag{3.35}$$
- **Voiced Chroma Accuracy**: probability that a truly voiced frame is correctly labeled as voiced and the right Chroma
$$\frac{TPCch}{GV} \tag{3.36}$$
- **Precision** [55], [56]: probability that a frame was correctly labeled both with voicing and pitch without taking into account wrongly labeled unvoiced frames.
$$\frac{TPC+TN}{TP+TN+FP} \tag{3.37}$$
- **Recall** [55], [56]: probability that a frame was correctly labeled both with voicing and pitch without taking into account wrongly labeled voiced frames
$$\frac{TPC+TN}{TP+TN+FN} \tag{3.38}$$
- **F-measure** [55], [56]: weighted average of the precision and recall.
$$\frac{2*precision*recall}{precision+recall} \tag{3.39}$$
- **Chroma Precision** [55], [56]**:** probability that a frame was correctly labeled both with voicing and Chroma without taking into account wrongly labeled unvoiced frames
$$\frac{TPCch+TN}{TP+TN+FP} \tag{3.40}$$
- **Chroma Recall** [55], [56]: probability that a frame was correctly labeled both with voicing and Chroma without taking into account wrongly labeled unvoiced frames
$$\frac{TPCch+TN}{TN+TP+FN} \tag{3.41}$$

- **Chroma F-measure** [55], [56]: weighted average of the Chroma precision and Chroma recall.

$$\frac{2 * chroma\_precision * chroma\_recall}{chroma\_precision + chroma\_recall} \tag{3.42}$$

4.4 Reverberation dataset: AIR

In order to obtain reverberated versions of the clean audio from the MIREX dataset different strategies were considered:

- Record audio reverberated: reproduce the clean audio with a speaker in different environments (e.g. rooms) and record the audio reverberated with a microphone. This option was quickly discarded when the MIREX was selected as the dataset; it would have been endless to record 1,000 excerpts in different scenarios. Besides, this process carries a significant amount of difficulties: speaker response, microphones distance and position, microphone response, etc. that need to be taken into account.
- Simulate spaces: use artificial impulse responses to convolve the clean audio or some kind of room/reverb simulator. This method is better than the previous one but results depend on a simulation or a specific model of reverberation.
- Convolve with real reverbs. This implies recording the impulse responses of different rooms and then using them to reverberate the clean audios. This option has the addition that you can study how the different properties of the impulse response (e.g. reverberation time) affect the results, but it is a long and meticulous process to obtain those impulse responses precisely. Luckily, the Institue of Communication Systems and Data Processing (IND) from Aachen, Germany, had already created a dataset of impulse responses perfectly suitable for our purposes, the Aachen Impulse Response dataset (AIR) [57], [58], [59] and that is what we used.

The Aachen Impulse Response (AIR) dataset is a set of impulse responses that were measured in a wide variety of rooms. The aim of the AIR dataset is to allow for realistic studies of signal processing algorithms in reverberant environments with a special focus on hearing aids applications. That is why it comes with binaural room impulse responses

(BRIRs) measured with a dummy head in different locations with different acoustical properties, such as reverberation time and room volume. It also provides impulse responses for hands-free reference point (HFRP) and hand-held position (HHP) (see Figure 4.3).

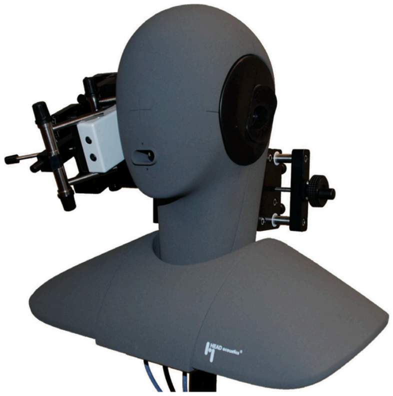

Figure 4.3 Dummy-head used in AIR dataset with the two-microphone mock-up clamped in the hand-held positioner.

In this study were always used monophonic impulse responses (using only one channel of the two provided) since we are not studying reverberation for our perception (with head-shadowing, etc.) but for machines. Thus, in all the selected rooms a configuration without dummy-head was used from the different distances available. Moreover, it was considered interesting to include both HHP and HFRP because are scenarios directly linked with mobile devices. Dummy-head results were computed and are available in the results appendix but have not been used in this study.

In

Table 4.2 is found the selection of rooms from the AIR dataset for this study as well as some data related to them. To further extend the information on rooms' characteristics see [57], [58], [59].

Rooms	Dimensions	Distance	RT60	Abbreviation name
Studio booth	3 x 1.8 x 2.2 m	0.5 m	80 ms	booth_1
		1 m	110 ms	booth_2
		1.5 m	180 ms	booth_3
Meeting room	8 x 5 x 3.1 m	1.45 m	210 ms	meeting_1
		1.70 m	220 ms	meeting_2
		2.25 m	240 ms	meeting_3
		2.80 m	250 ms	meeting_4
		HHP	254 ms*	meeting_hhp
		HFRP	337 ms*	meeting_hfrp
Lecture room	10.8 x 10.9 x 3.15 m	2.25 m	700 ms	lecture_1
		4 m	720 ms	lecture_2
		5.56 m	790 ms	lecture_3
		7.1 m	800 ms	lecture_4
		8.68 m	810 ms	lecture_5
		10.2 m	830 ms	lecture_6
		HHP	236 ms*	lecture_hhp
		HFRP	818 ms*	lecture_hfrp
Office room	5 x 6.4 x 2.9 m	1 m	370 ms	office_1
		2 m	440 ms	office_2
		3 m	480 ms	office_3
		HHP	390 ms	office_hhp
		HFRP	520 ms	office_hfrp

Table 4.2 Rooms selected from the AIR dataset. Values with * were calculated directly from the impulse response using the Schroder method [60] because the original AIR papers [57], [58] did not include that information.

4.5 F0 estimation algorithms

Fundamental frequency estimation algorithms were selected taking into account, in one hand, their current relevance in the field of F0 estimation (i.e. the most used and accurate) and, in the other hand, those specially focused towards the singing voice (or speech) and with background interferences overcoming capabilities.

The fundamental frequency estimation algorithms selected were:

- **Two Way Mismatch (TWM)**[16]: it is a frequency-domain algorithm based on harmonic pattern matching (see Section 2.2.2.c). It is widely used nowadays for its accuracy and efficiency. It was used a proprietary implementation from the MTG that could be fine-tuned to adapt to the singing voice.
- **Spectral Auto-Correlation (SAC)**[13]: it is a frequency-domain algorithm base on spectrum auto-correlation (see Section 2.2.2.b). It is interesting for its special accuracy and robustness. As it happened with TWM, it was used a proprietary implementation from the MTG that could be fine-tuned to adapt to the singing voice.
- **YIN**[6]: it is a modification of the time-domain auto-correlation algorithm (see Section 2.2.1.b). It is one of the currently most used algorithms, especially for voice because it was specially conceived for speech. Among its virtues excels its speed (it is the fastest of the four algorithms used here), being able to run in real-time needing only to buffer twice the period to detect (other methods usually require a bare minimum of four periods). It was used the original implementation from the author described in [6].
- **MELODIA**[21]: it is a frequency-domain algorithm based on harmonic summation to calculate a salience function. It is different than the previous ones because it is especially designed to perform in a multi-pitch environment, i.e. it extracts the main melody of songs that have background music. This is exactly what makes it interesting for our study because this inherent capacity to overcome background interferences can make it more robust facing reverberation. It was used the original Vamp plugin implementation from the author described in [21]. A disadvantage of MELODIA is that its methodology can only be applied offline, i.e. not in real-time.

4.5.1. Algorithm fine-tuning

In order to compare the performance of the algorithms in the best conditions its execution parameters were carefully selected. In one hand, all algorithms had to use a fundamental frequency search range of 63.1883 to 706.3069 Hz. It is exactly the range that the ground truth covers plus a half tone in both extremes. This half tone is added so algorithms still fail on the extremes (remember that our evaluation considers a pitch correct if it is closer than a quarter tone). Besides, all algorithms used the same hop size (20ms) and window size (40ms) than the ones used for the MIR-1K ground truth. Nevertheless, it had to be made an exception with MELODIA because window and hop size are fixed and cannot be changed. Because its the fixed values were smaller, it was safely used linear interpolation. On the other hand, the policy to follow regarding the other parameters, which are different in every algorithm, was to keep them with the default value as long as it was reasonably optimum for our task. Besides, for optimization it was always considered a scenario with clean audio (no reverberation) because the goal of the study is to observe the performance degradation of the algorithms when set to properly work in standard conditions with singing voice, but not to see which of them performs better when best optimized with reverberated audio. That being said, the rest of the parameters were set to the default value with the following exceptions:

- TWM and SAC were actually set to the default parameters adequate for voice, i.e. the voice profile.
- In YIN existed a doubt about the parameter that decides the voicing of a frame. YIN outputs, along with the F0 contour, an aperiodicity measure (AP0) that is used to detect the voicing. When a frame is substantially aperiodic, i.e. it is above a threshold (0.2 by default), it is considered unvoiced, or voiced otherwise. To prove the default value was reasonably optimum there were generated two histograms. The first one was an histogram of the aperiodicity measure of all voiced frames in MIR-1K (see Figure 4.4), and the second one the same for unvoiced frames (see Figure 4.5). It is easy to see on those histograms how this default threshold of 0.2 is correctly placed to discriminate the voicing of a frame. It could have been used a more complex strategy (e.g. grid-search) to find exactly the most optimum threshold but, since the goal was only to corroborate the

reasonable suitability of the parameter value for our task, it was considered enough.

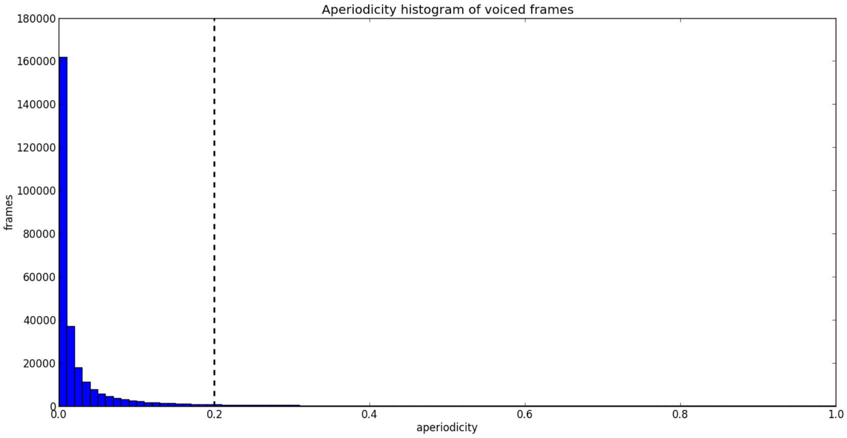

Figure 4.4 Aperiodicity histogram of voiced frames

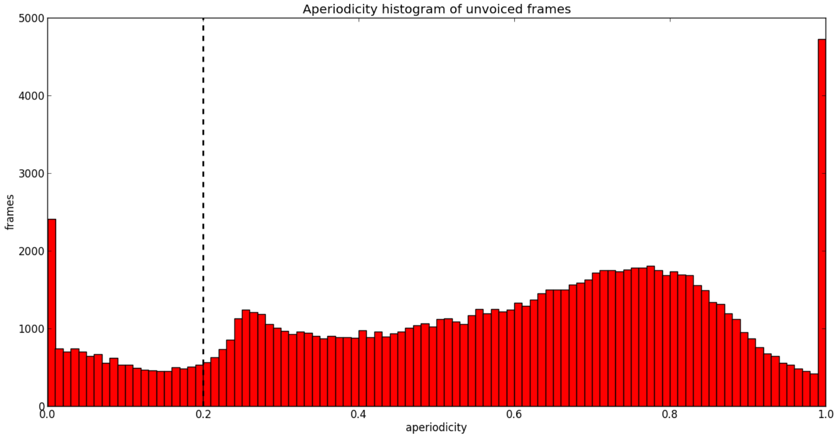

Figure 4.5 Aperiodicity histogram of unvoiced frames

- MELODIA offered two extra parameters that needed to be properly set to work with monophonic audio: voicing tolerance, a threshold to discard weak salience contours, and monophonic noise filter, a voicing filter to label as unvoiced frames with low loudness. It was performed a grid search using the MIR-1K-Micro dataset evaluating 10 steps for voicing tolerance (also called minpeaksalience) and 14 steps for monophonic noise filter (also called voicing). Results, which can be seen in Figure 4.6, show how the best accuracy is obtained with the maximum of both values (3 for monophonic voice filter and 100 for voicing tolerance). It is interesting to observe that the monophonic voice filter has a big influence in the final accuracy while voicing tolerance has practically no effect.

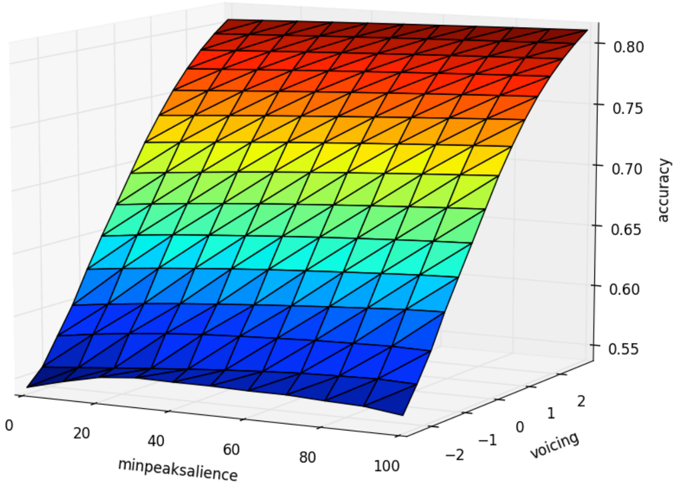

Figure 4.6 Results of the grid-search (3D representation) for MELODIA

4.6 De-reverberation algorithms

The last task was to de-reverberate already reverberated signals to see if the robustness of F0 estimation could be improved and, at the same time, see if this de-reverberation had any negative effects if applied on non-reverberated signals. The de-reverberation methods for this study needed to be blind, i.e. with no knowledge of the reverberation or its properties, and focused on the singing voice. Indeed, speech focused de-reverberation techniques were considered since there are none specifically for the singing voice. There exist many proposals for de-reverberation given these constraints but it is actually an unresolved problem in blind conditions and there are few implementations. Finally, all the (very few) publicly available implementations were considered. It is important to point out that, in a similar fashion than what was done with the F0 estimation algorithms, the de-reverberation algorithms were used with the default parameter values, adapting only to the sample rate in case of needed.

4.6.1. NMF de-reverberation

NMF de-reverberation [47] is a method based on linear filtering (see Section 3.2.1) in the spectral domain. It considers a different and time-varying impulse response filter for every gamma-tone sub-band (gamma-tone sub-bands are a spectral representation useful to emulate human perception [61]). It then de-convolves reverberated speech into clean speech and the impulse response filter using Non-negative Matrix Factorization (NMF). NMF is constraint to the non-negativity of the spectral amplitude values and the sparsity of the speech matrix. Finally, the difference between the magnitude of the observed signal and the estimated signal is minimized using the mean-squared error. Figure 4.7 shows a schematic overview of the process.

This method was conceived to improve Automatic Speech Recognition (ASR) results in reverberant environments. Probably because of this (it has a different target than F0 estimators) and because of the weak constraints imposed in the NMF decomposition it did not improve our results but got them lightly worst. This is why this method is not presented in our results.

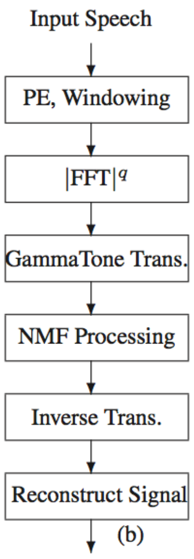

Figure 4.7 NMF de-reverberation processing

4.6.2. ITD de-reverberation

ITD de-reverberation [48] is a modification on NMF de-reverberation and, thus, has the same basic processing steps (see). The difference is that when de-convolving the signal it uses the NMF decomposition as an initialization step and then uses an iterative process to improve this deconvolution using a constraint of non-negativity of the speech spectra magnitude.

Even though this method performs slightly better than NMF de-reverberation in our results it has the same weaknesses, thus it does give any improvement in our results and consequently it is neither present in our results. Besides, the iterative process supposed a drastic increase in computation time, so results were only calculated for the MIR-1K-Micro dataset.

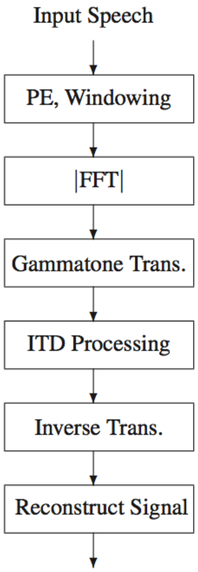

Figure 4.8 ITD de-reverberation processing

4.6.3. NML de-reverberation

NML de-reverberation [62] is a method based on spectrum enhancement (see Section 3.2.2). It models the signal as a convolution in time, separating the direct-path of the signal, early reflections and late reflections:

$$x(n) = h(0)s(n) + \sum_{i=1}^{\tau} h(i)s(n-i) + \sum_{i=\tau+1}^{T-1} h(i)s(n-i) \qquad (3.43)$$

where $x(n)$ is the captured signal (with reverberation), $s(n)$ is the clean signal without reverberation, $h(n)$ is the impulse response of the reverberation, T is the duration of $h(n)$ and τ is a number between T and 0 (typically around 50ms) that separates the early and late reflections. This method then applies a multi-step linear predictor [63] that predicts the coefficients $w(n)$ of the late reflections:

$$x(n) = \sum_{p=0}^{N-1} w(p)x(n-p-\tau) + e(n) \qquad (3.44)$$

where N is the number of coefficients of the multi-step prediction as well as the number of points of the predicted late reflections impulse response and $e(n)$ is the error to minimize using the mean-squared error minimization process. Then these coefficients are used to obtain the actual late reflections by means of convolving the coefficients with the captured signal in the time domain. Finally, the late reflections are removed from the captured signal using spectral subtraction [64]. NML de-reverberation also applies a process of pre-whitening, using a short 20-steps linear predictor, in order to de-correlate early and late reflections. A schema of the process is depicted in Figure 4.9.

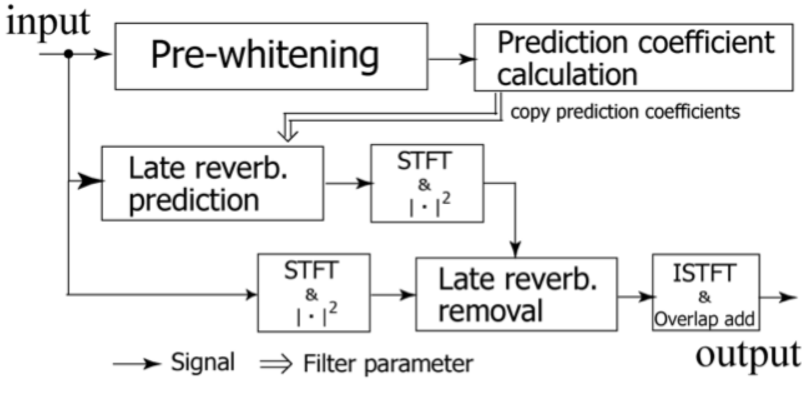

Figure 4.9 NML de-reverberation processing

NML de-reverberation has been used for ASR improvement [42], [65], [66] as well as reverb reduction for studio production [62]. Even though there is not a public implementation of the system it exists an RTAS (Real Time AudioSuite) implementation, i.e. an audio plugin to be used with the industry standard DAW Pro-Tools. This supposed that every single file had to be processed one by one and, thus, the results were only computed for the MIR-1K-Micro dataset. This is the only tested de-reverberation algorithm that improved our results and it is analyzed in Section 5.3.

5 Results analysis

In this section the results of the study are depicted and analyzed to extract conclusions. To understand the following subsections is essential to know it was proposed, as an improvement, a new fundamental frequency algorithm that is called MIX in this dissertation. This method, instead of being a complete new algorithm, is just a proof of concept on how the results of this study can be used to improve the state of the art F0 estimators to become more robust in reverberant conditions. MIX uses the voicing detection mechanism of MELODIA and the pitch detection algorithm of SAC using the NML de-reverberation as a preprocessor (only for SAC). It is a mix (hence its name) of the best performing parts of the methods evaluated in this study.

5.1 Dry conditions

In this section are analyzed the results of the all the rooms selected from the MIR-1K dataset (including HHP and HRFP, see

Table 4.2 in page 59) in dry conditions, i.e. using the dataset as it is. These results are the average of the individual results of every excerpt; being these individual results the average of all the frames of the excerpt. The overall results in dry conditions are depicted in Figure 5.1.

The best performing algorithms in accuracy are YIN, MELODIA and MIX, shortly followed by SAC. This classification can be corroborated with the f-measure (an evaluator also widely used), where the results are in general slightly higher but the classification remain practically identic. These good results, above 90% of accuracy, prove the validity of both MIR-1K dataset, despite its recording interferences, and selected F0 estimators. Those overall accuracies are achieved in slightly different ways. YIN is obtaining the best voicing accuracy thanks to the good compromise between its voicing false alarm and voicing detection, but is being overpassed by all but MELODIA in pitch accuracy. MELODIA is getting much of its score thanks to its good voicing detection; observe how it has a practically 0% voicing false alarm (less is better). SAC is performing the best in terms of pitch detection and has a decent voicing mechanism. TWM is obtaining not very good results, with 77% of accuracy, but this is not because of its pitch detection, which is very good (95%), but because it has a faulty voicing mechanism; it is considering voiced most of the frames (it has a very high voicing false alarm). It is important to note how MIX, that is our final proposal for F0 estimation in reverberant conditions, performs as good as the best algorithms in dry conditions. This makes this algorithm ideal for situations where noticeable reverberation can be captured or not, for example mobile devices, without having to worry about reverberation presence. It can be seen that raw pitch accuracy of MIX is slightly lower than SAC and they are supposed to be the same. This is due to the NML de-reverberation method shortening some of the excerpts that finish with voiced frames.

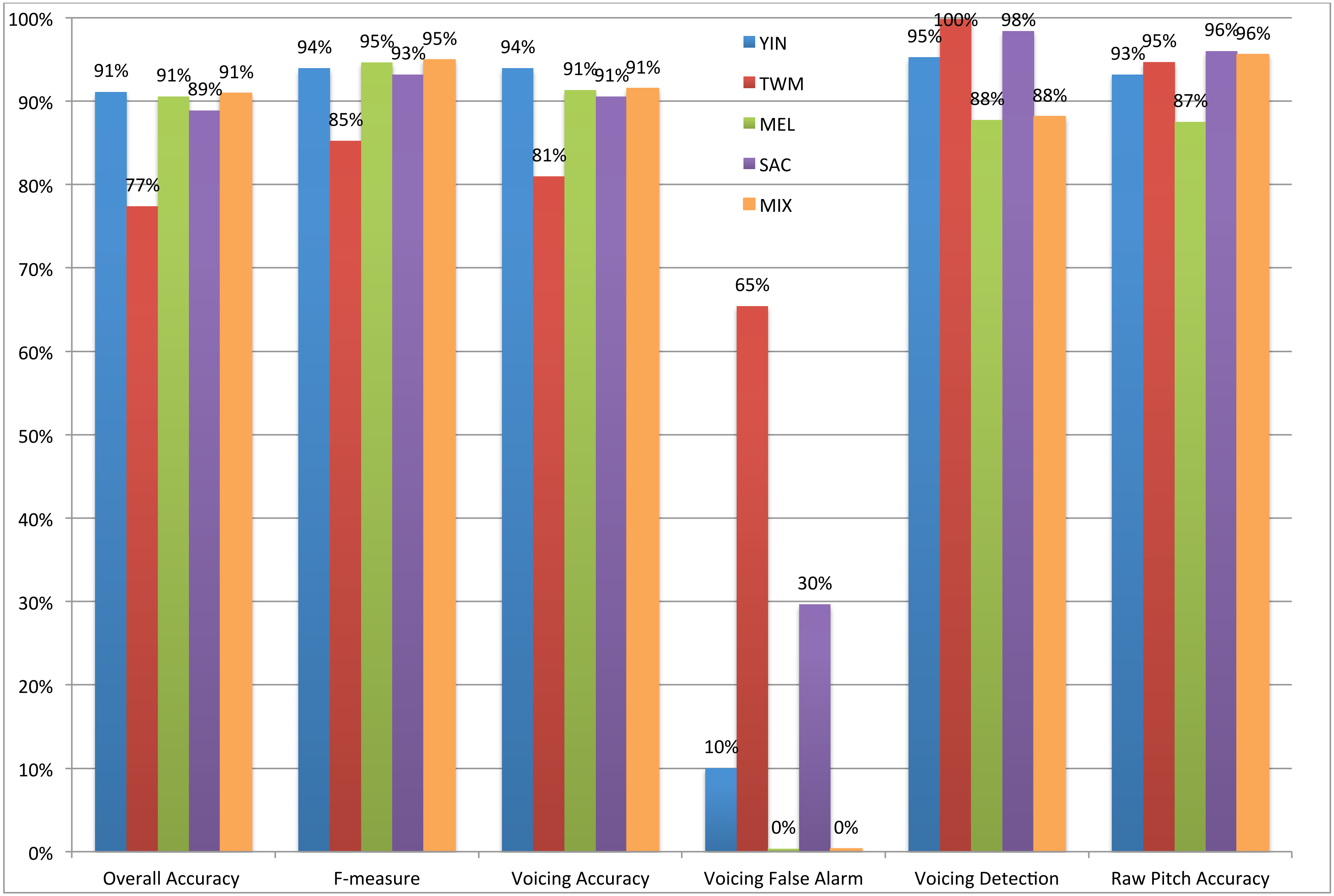

Figure 5.1 Overall results in dry conditions

5.2 Reverberant conditions

In this section are analyzed the results obtained after applying reverberation (using the AIR dataset impulse responses). These results are analyzed in function of the reverberation time (RT60) taking into account a higher reverberation time produces, in general, more degradation. To see the correspondence between the figures' points and the rooms use

Table 4.2 in page 59. It is important to see that in the figures the points corresponding to the same room are clustered together, so rooms appear in the following order (from less to more RT60): studio booth, meeting room, office room and lecture room. This fact explains the "jumps" (i.e. change in tendency) between some points that are due to a change in room. Moreover, in this section will be only studied the degradation produced in the already existing F0 algorithms and, thus, MIX is no considered.

The first results to analyze are the overall accuracy, depicted in Figure 5.2. The first thing to corroborate is the negative effects that reverberation produce for all F0 estimators. YIN performs the best in dry conditions but it degrades the most in presence of reverberation. TWM and SAC degrade similarly and MELODIA is the more resilient. The main conclusion to extract here is that YIN is a bad choice for reverberated signals and MELODIA is the best choice but, anyway, there is no F0 estimator capable of preserving acceptable accuracies when surpassing reverberation times higher than 200ms. To analyze more profoundly why this is happening and how can it be improved the results are going to be analyzed with pitch and voicing isolated.

Raw pitch accuracies can be seen in Figure 5.3, those accuracies refer to how good the algorithm is capable of detecting the pitch without considering its voicing mechanism. The first noticeable issue is that, although MELODIA was the algorithm with the best accuracy, SAC is the best performing algorithm with pitch. MELODIA is still the algorithm that degrades the least but SAC has better performance for all reverberations. TWM is close to be as good as SAC in its pitch detection. YIN, instead, degrades very fast. In a few words, using the pitch detection mechanism of SAC would be the best choice.

Voicing accuracies can be seen in Figure 5.4, those accuracies refer only to the voicing mechanism of the algorithms with no consideration for the pitch detection. MELODIA degrades the least and with a huge difference in contrast with its competitors. Is this very good voicing mechanism that is giving it the best overall accuracies to MELODIA. YIN is still better only for short reverberation times that are less than 100ms long. This is the first clue to observe that there is some kind of threshold between 100 and 200ms of RT for which algorithms do not degrade much. Nevertheless, YIN is degrading vertiginously faster than the other algorithms, so its voicing detection mechanism is in any way robust to reverberation. Here SAC is close to be as bad as YIN. TWM, in the other side, seems not to degrade much but it has to be considered that it already starts from a very low value compared with the other algorithms and, thus, it is far more complicated to get worse when you are already bad.

In conclusion, it was observed that even though MELODIA has the best overall accuracy it is very good with voicing but not that much with pitch detection, where SAC outperforms it.

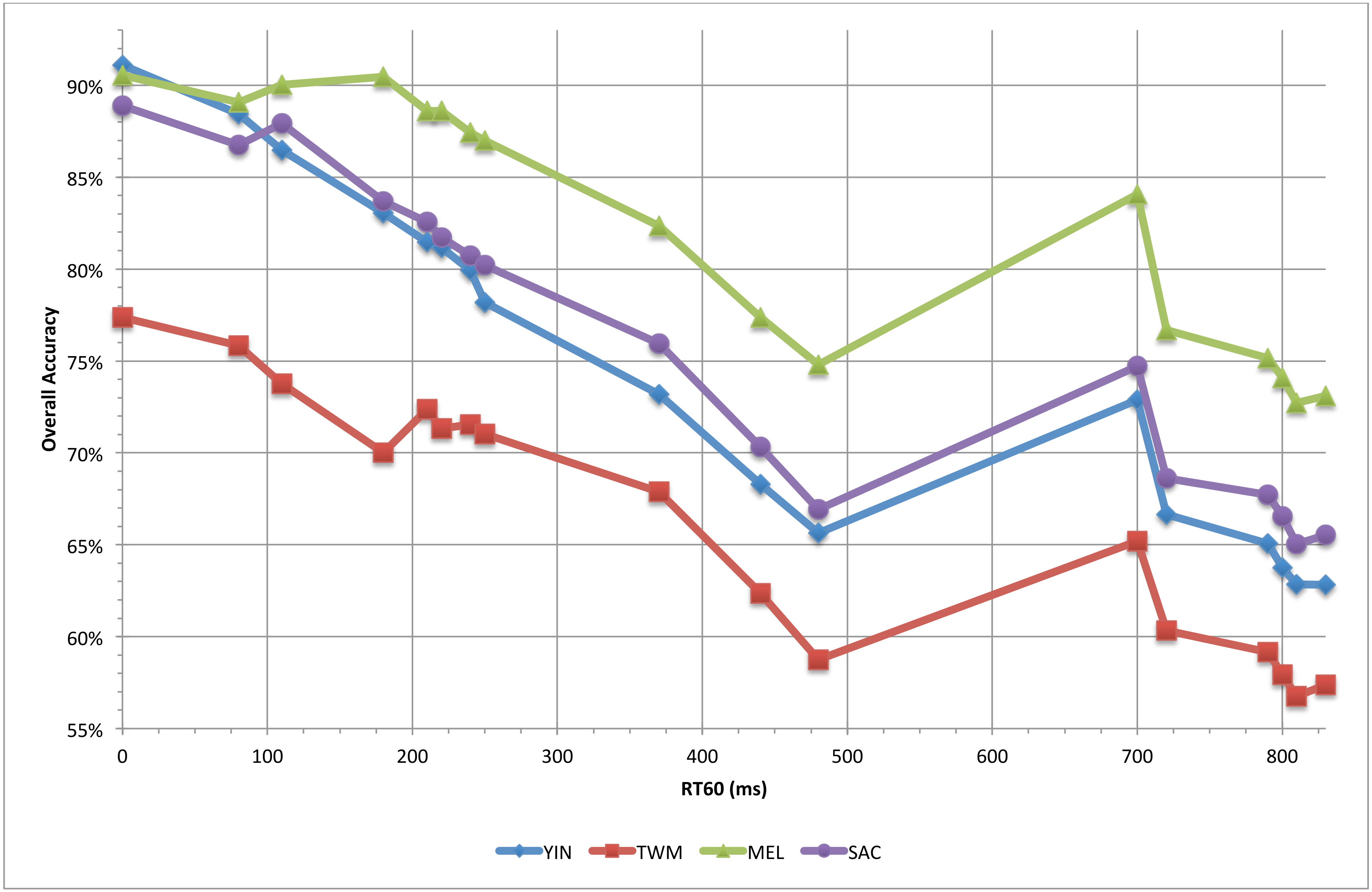

Figure 5.2 Overall accuracy in reverberant conditions

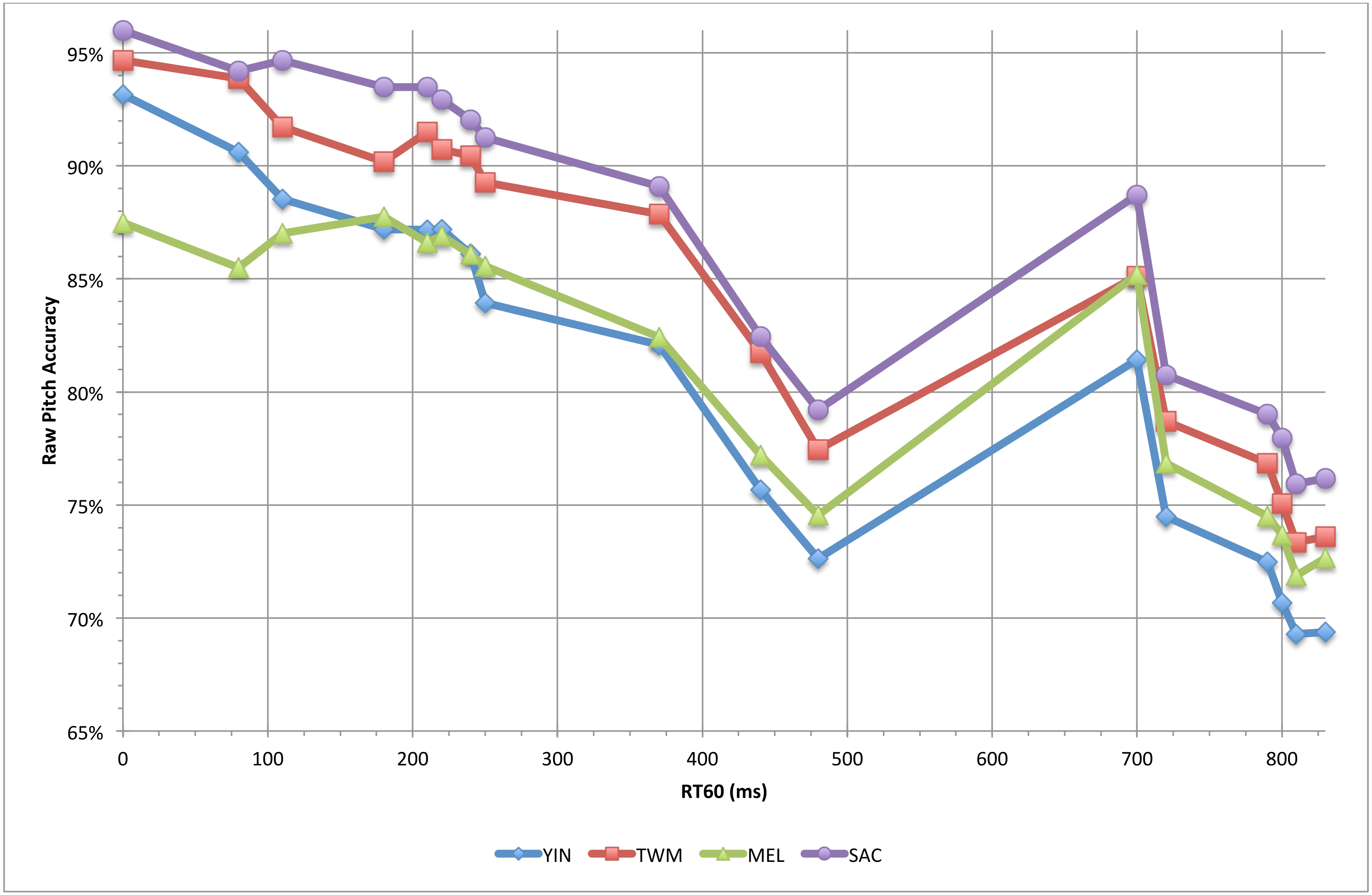

Figure 5.3 Raw pitch accuracy in reverberant conditions

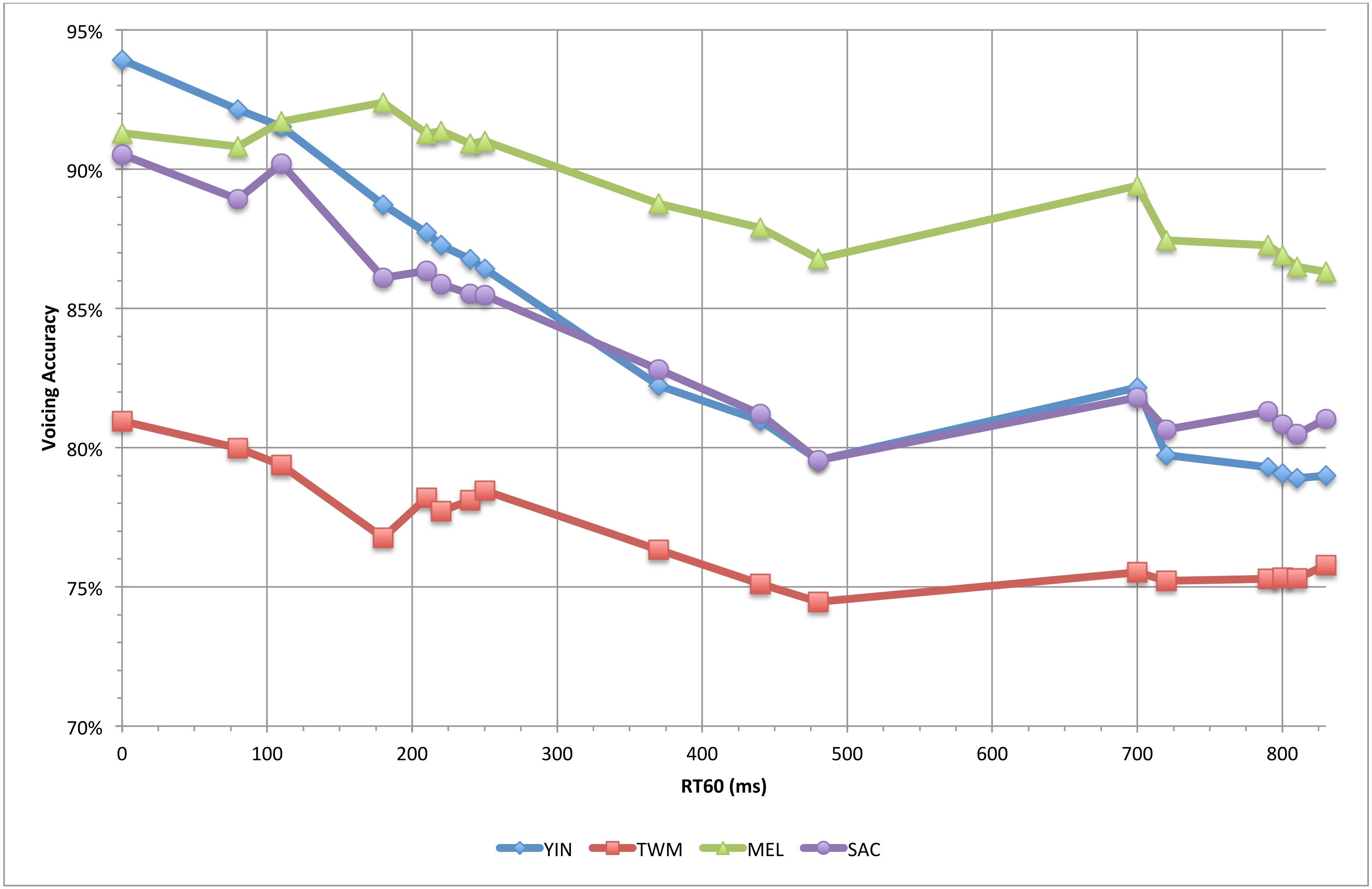

Figure 5.4 Voicing accuracy in reverberant conditions

5.3 Using de-reverberation

In this section the results analyzed make use of the NML de-reverberation method. Even though we evaluated three de-reverberation methods (see Section 4.6) this was the only one that improved results. The graphics follow the same disposition than in the previous section (read Section 5.2 for more information). Indeed, there will be presented the same graphics with the addition of the same algorithms with NML as a preprocessor. For example, YIN is the solid blue line and YIN+NML is the dotted blue line. MIX is also included to observe the final results of our proposed estimator.

In Figure 5.5 can be seen the overall accuracy. The first thing to notice is that NML improves the accuracy of all algorithms with the exception of MELODIA, which gets notably worst. Despite of that, MELODIA by itself (without NML) is still the best performing algorithm of the originally selected ones in this study. Only SAC+NML gets close to MELODIA performance. The improvement experienced in YIN is subtle while for TWM is big, getting to overpass the original YIN. Nevertheless, it can be seen how our estimator MIX is the best of all estimators, being SAC+NML superior only with RTs inferior to 100ms. This confirms again the existence of a threshold for short RTs to be treated differently. It is also interesting to observe that with dry conditions (RT=0ms) all algorithms but MELODIA also improve. Noticeable is the improvement of SAC and TWM.

Regarding the raw pitch accuracy, depicted in Figure 5.6, all algorithms improve but MELODIA, again. It seems very clear, due to how much worst MELODIA gets, that NML is not helping it in the voicing mechanism. The best performing algorithm is SAC+NML being closely followed by TWM+NML, which is the only one to improve with NML in dry conditions. It is interesting to point out that here also can be seen a threshold for short reverberation times where NML starts improving results instead of getting them worst. This totally makes sense since applying a de-reverberation method to a dry or close-to-dry signal should only get it worst or untouched at least. Eventually, from this graphic can be concluded that, although MELODIA has the best overall accuracy, it can be significantly improved using SAC+NML (even SAC alone). In this figures MIX is not plotted since it is exactly the same as SAC+NML.

In Figure 5.7 is shown the voicing accuracy, where MELODIA without NML continues to be the best performing one. Once again, all algorithms improve but MELODIA with the use of NML. It is noticeable how we can find the short RT threshold present here also.

This time it can be observed that below 200ms there are better performing algorithms than MELODIA, for example SAC+NML. Besides, SAC and TWM are improving with NML in dry conditions. It has to be pointed out that in this graphic MIX has been plotted because, although it is using the same voicing detection mechanism than MELODIA, it does not have the same exact results. This is because when SAC labels a frame as unvoiced but MELODIA says it is actually voiced we do not have the pitch information (SAC does not provide it) so the frame must remain as unvoiced.

In conclusion, it can be seen how NML de-reverberation improved the performance of all methods except MELODIA that, because of its different target of working with background interferences, does not fit with NML as a preprocessor. It is also observable that NML generally improves results for reverberation times longer than 100-200ms but not always with shorter ones, where there seems to be a different behavior. This is possibly because short reverberation times do not interfere enough the algorithm until a certain reverberation time. Something probably linked with the window length the algorithms use. In this case, a simple solution would be to treat reverberation as a common filter and use Cepstrum Mean Normalization (CMN) on a frame basis to reduce it. Another explanation would be that short RTs belong to the studio booth, that has a flat frequency response opposed to the other rooms and, thus, produces different and less harming interferences. Another possibility would be that as NML is using fixed parameters they expect a certain reverberation length and consequently do not handle properly too short and too large RTs.

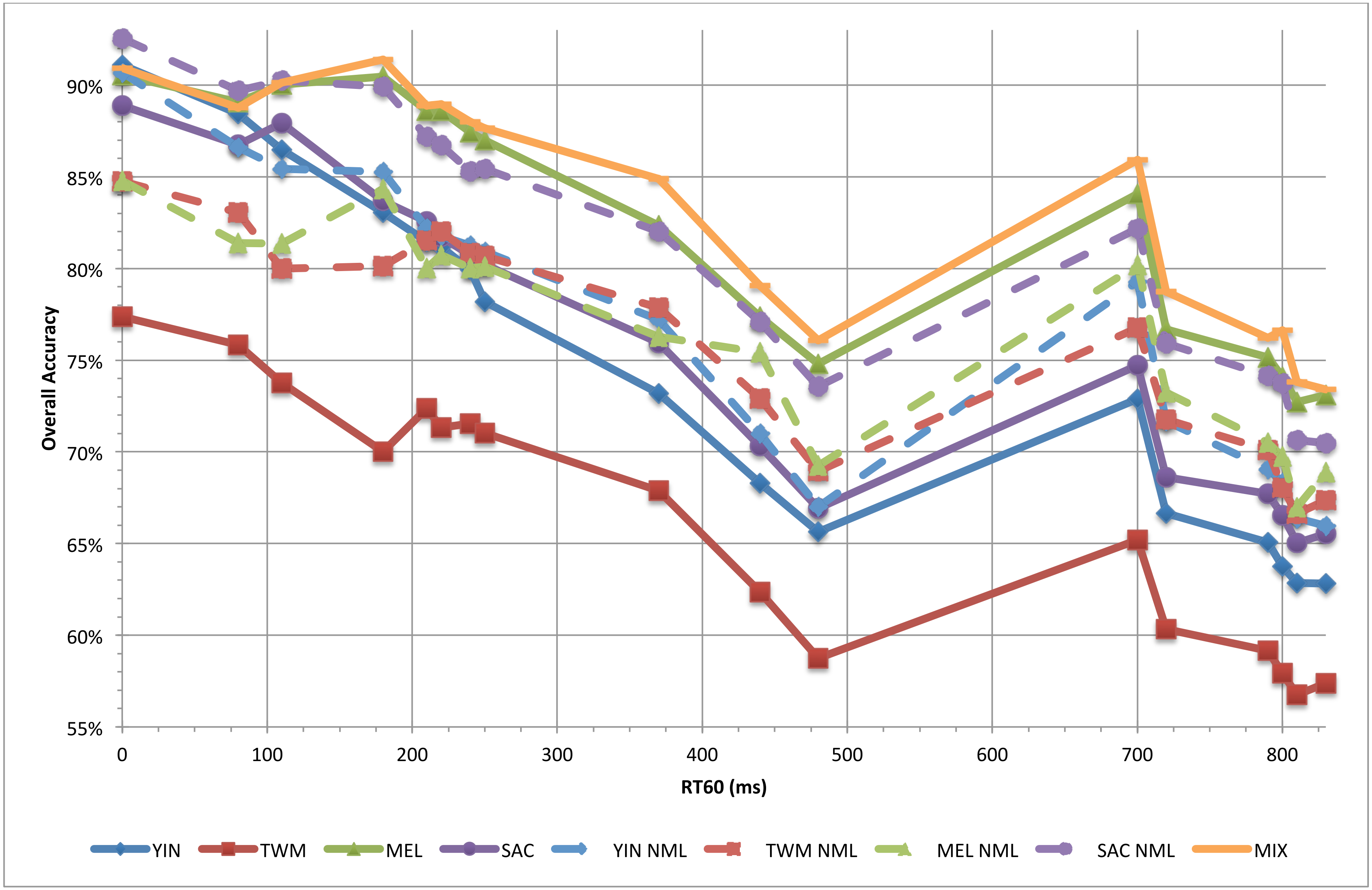

Figure 5.5 Overall accuracy using NML

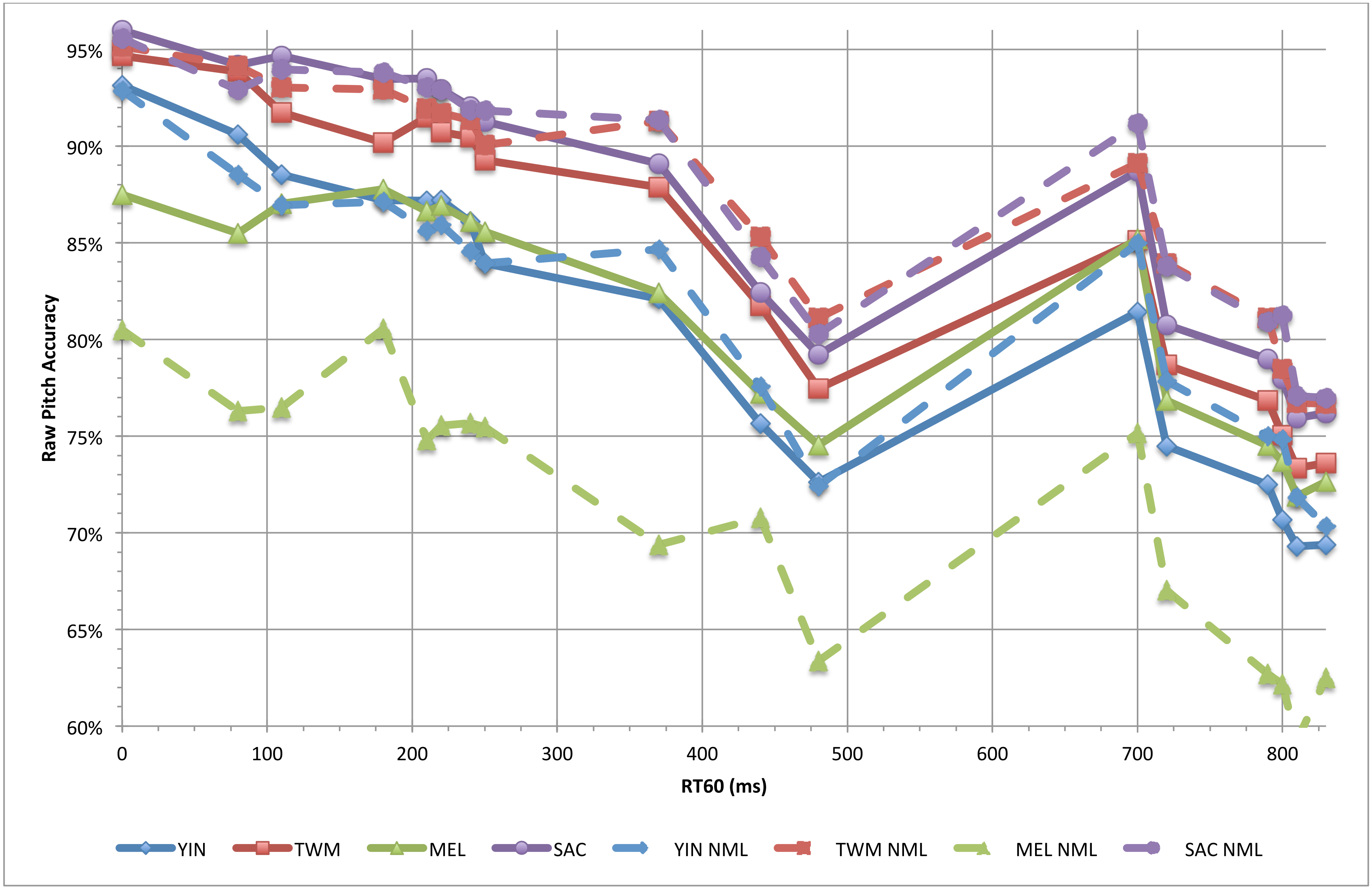

Figure 5.6 Raw pitch accuracy using NML

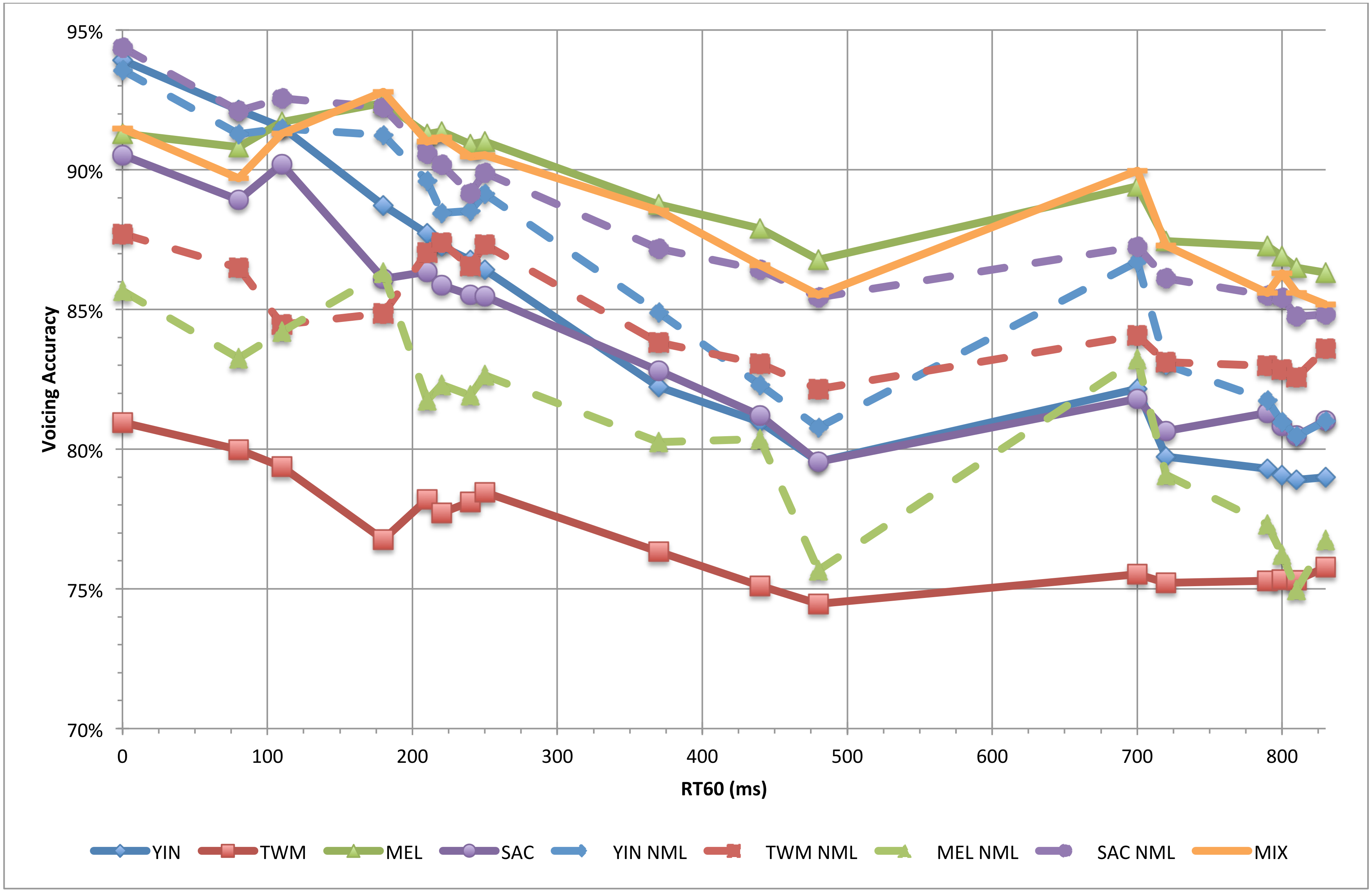

Figure 5.7 Voicing accuracy using NML

6 Conclusions

In this section the final conclusions of this study are presented. For partial conclusions or to see where the conclusions here come from please refer to Section 5 Results analysis.

This thesis is a study to observe the weaknesses of fundamental frequency estimation algorithms with reverberated signals and its goal is to find the most robust ones. First of all, it has been found that F0 estimators behave similarly than human perception in the sense that short reverberations do not affect them much (in humans it even helps) but long reverberations affect its performance. Nevertheless, humans still accept larger amounts of reverberation than those that machines can handle and, thus, it is a field with much research to be still carried on.

There were tested four F0 estimators focused towards the singing voice: YIN, TWM, SAC and MELODIA. It was shown that the best performing one was MELODIA. This best response is due to the special design of MELODIA; prepared to extract the main melody of a song with background music. Is this capability of discarding background interferences that made it outstand. Anyway, a deeper insight showed this discarding capacity was very good in terms of voicing decision but not for pitch detection, where it was surpassed by SAC and TWM even under reverberation conditions.

Later, in order to improve the state of the art of F0 extraction methodologies we applied three different de-reverberation techniques as preprocessors. Only one of them, NML, could improve the previous results. The other two, NMF and ITD were not useful for our task because they where too focused on just automatic speech recognition (ASR) and they where using a non-negative matrix factorization to de-convolve clean signal and reverberation with too weak constrains. NML, in the other hand, besides of having proved its value in improving ASR algorithms in the past, is being used for studio production de-reverberation. Indeed, this is the implementation that was used here. NML improved results for all algorithms except MELODIA, which did not fit with a preprocessor because it is already focused on handling with interferences. Despite of the improvement NML gave on the other algorithms, MELODIA still performed generally the best. Anyway, we continued observing that the pitch decision mechanism of MELODIA could be improved. So, as a proof of concept on how the results of this study can be used to advance on the field, the MIX fundamental frequency estimator was ideated. MIX combines the best performing parts of this study, i.e. the voicing mechanism of MELODIA and the pitch detector of

SAC using NML as a preprocessor. With this combination it was possible to get an average improvement of 2% comparing with MELODIA results (the best ones) and, very importantly, it performed as good as the best F0 estimators (91% accuracy) in dry conditions, i.e. with no reverberation.

In other words, MIX produced the best results when estimating the fundamental frequency of reverberated signals. As MIX is not actually implemented but just a proof of concept the best performing F0 estimator that can be used “as it is” is MELODIA. De-reverberation techniques are still a work in progress but can be used as a preprocessor of standard F0 estimators to improve its results. It is needed to take into account that not any de-reverberation method will be suited for such a task and more research is needed to be done in this direction. Nonetheless, MIX had a 25% less accuracy for the longest reverberations using recordings a human can still comfortably understand and this indicates there is still a lot of margin to improve F0 estimators to be more robust to de-reverberation.

Finally, it would be interesting to think out of box. This dissertation is completely signal oriented. But we should not forget that we are trying to solve a concrete problem: overcome the detrimental effects of reverberation for machines, specifically for fundamental frequency estimators. Nowadays, it is becoming more common in the audio computing field to understand that most of the unsolved problems cannot be sorted out with just the signal processing but adding context information. It would be an enriching activity that some whiles we stop our mechanical solving strategies to approach things in a different way. We, as humans, posses a lot of extra information (e.g. source distance, room dimensions, temperature and a very large etc.) that machines do not, and that could be very useful to solve our research problems. To point out a simple example, our mobile phones have a light sensor that change screen luminosity automatically depending on the darkness they perceive. Another example is that some mobile phones have a second microphone to cancel background noise. Then, it would not be that rare to see mobile devices in some years with ultrasound sensors capable of emitting ultrasounds to perceive an approximate of the room dimensions surrounding it. That would help, for example, location, automatic sound level control and, why not, reverberation cancellation. That might not be the best example but it is useful to understand that sometimes signal processing approaches cannot just go further. In this dissertation it was believed and then corroborated that there is still work to do in making machines understand reverberation through signal processing, but would that be enough?

7 Future work

This section proposes different tasks to accomplish in order to extend this study and/or progress in the consecution of a reverberation robust fundamental frequency estimation algorithm.

- This study has focused on finding the most robust F0 estimation algorithm when the estimators where set to work properly with dry signals. It would an evident next step to fine-tune the algorithms parameters to perform the best in reverberant conditions.
- In this dissertation it has been shown that NML de-reverberation method improves F0 estimators performance in reverberant conditions. NML was set to work with default parameters. It would be interesting to explore the possibility to adapt NML parameters depending on the input signal. More specifically, NML has to main parameters: late reflections length and late reflections delay (i.e. how far in time are late reflection from the direct path), it makes sense to adapt these parameters if we have some previous (even if approximate) knowledge of the reverb characteristics; e.g. use some reverberation time estimator. Indeed, NML uses a multi-step linear predictor that is used to obtain late reflections. This late reflections could be analyzed to recalculate iteratively the multi-step linear predictor until some convergence criteria. Besides, this approximate information about the amount of reverberation in the signal could be used to change the parameters of the F0 estimator in real-time. Moreover, this same information could be used to change the strategy (F0 algorithm, NML on or off, etc.) taking into account the threshold detected in our results analysis (Section 5, RT of 100-200ms).
- Only one out of three de-reverberation methods succeeded in improving F0 estimators' results. We could not try more de-reverberation methods mainly due to lack of them being publicly available. An effort on sharing more openly the de-reverberation methods proposed is needed and of general interest, specially considering that this field still needs lots of research. In our particular case, testing more de-reverberation methods would be essential to find what do they have in common that helps F0 estimators. Having this information, a special strategy

could be integrated in the F0 estimator eliminating, then, de-reverberation preprocessing and, thus, reducing computing overhead.

- In this study it has been used the MIREX audio melody extraction evaluation in order to use the currently most extended evaluation method. Despite of that, depending on the final application that the F0 estimators will be used on this evaluation could not be appropriate and so, a different evaluation would be needed. A clear example would be that, while our evaluation method considered a pitch correct with a maximum deviation of a quarter of tone, many applications need more precision.
- MIR-1K has been very useful for our study, it is very large and the results obtained from it have en extra of confidence for this reason. But it has also been observed some weaknesses in it (see Section 4.2) as recording interferences and having used mostly amateur signers. It would be useful to compare results using studio standard quality recordings and professional singers.
- This dissertation was focused on monophonic signals since it is the general scenario when capturing the singing voice. On the other hand, nowadays, it is becoming more common to have multi-channel recordings (e.g. most new mobile devices integrate two microphones), that could be used both to use multi-channel de-reverberation methods and multi-channel F0 estimators that take advantage of this extra information (spatial cues, sound location, etc.).

Results appendix

In this section the numerical results of the study are included. Because the dataset used, MIR-1K, has so many excerpts (1,000) we needed to create a smaller subset of it, MIR-1K-Micro; see Section 4.2 for more details on its creation and an explanation on why it is acceptable to compare results from the different datasets. By default, **all results presented are extracted from the complete MIR-1K dataset** with the following **exceptions**:

- Results where **NML** de-reverberation has been used are computed with the MIR-1K-Micro dataset. Our implementation consisted on a RTAS plugin that needed to be executed excerpt by excerpt in the Pro-Tools DAW without possible automation.
- Results where **ITD** de-reverberation has been used are computed with the MIR-1K-Micro dataset. The iterative process that the ITD algorithm slows down it to the point it would have been technically endless to use the complete MIR-1K.

Numerical results are presented in tables. The first table contains dry results, with different F0 algorithms in columns and evaluators in rows. Then, following tables correspond to a single evaluator with different F0 algorithms in columns and different impulse responses in rows. An abbreviated name has been used for the impulse responses that can be found at

Table 4.2 (page 59). The abbreviation "dry" means it was used no impulse response just the dataset as it is. Regarding the F0 algorithms' name it has been added NML to the end of its name when NML de-reverberation was used as a preprocessor, e.g. YIN NML for YIN. Reverberation time (RT60) is expressed in milliseconds. All evaluators have a value range of [0,1] being 1 the best result and 0 the worst, except for the voicing false alarm that works the opposite way.

For more details on the different rooms and distances used to reverberate the signal see

Table 4.2 (page 59). To extend evaluators' information see Section 4.3.

Dry results

Results for dry excerpts, i.e. MIR-1K as it is. All the rooms from

Table 4.2 (page 59), including HHP and HFRP have been used.

	YIN	TWM	MEL	SAC	MIX
Overall Accuracy	0.91089378	0.773665165	0.905363007	0.888691586	0.909224376
Voicing Accuracy	0.9393083	0.809527809	0.913013025	0.905283868	0.914807908
Voicing False Alarm	0.100028113	0.654046238	0.003460538	0.296280634	0.003458967
Voicing Detection	0.952218869	0.99834949	0.877206184	0.983764904	0.881185376
Raw Pitch Accuracy	0.931407144	0.946640946	0.874813882	0.959897325	0.955573669
Voiced Pitch Accuracy	0.911834824	0.946640946	0.866175817	0.959897325	0.873487836
F-measure	0.939125742	0.852140022	0.946130905	0.931791842	0.94928794
Precision	0.941764283	0.774580449	0.990510253	0.898432415	0.99283092
Recall	0.936943412	0.953633674	0.90630817	0.96947688	0.910132145
Overall Chroma Accuracy	0.927714548	0.788869605	0.906587496	0.893395498	0.907836509
Raw Chroma Accuracy	0.969673454	0.969019864	0.897903243	0.968728003	0.968179727
Voiced Chroma Accuracy	0.935760189	0.968758965	0.868025045	0.966754144	0.871166822
Chroma F-measure	0.956546121	0.868981923	0.947415278	0.936768639	0.947836648
Chroma Precison	0.959305594	0.789807654	0.991857114	0.903231064	0.991312008
Chroma Recall	0.954254766	0.972609127	0.907535954	0.974653782	0.90873983

Overall accuracy

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.91089378	0.773665165	0.905363007	0.888691586	0.906762952	0.847343855	0.847772784	0.925338415	0.909224376
booth_1	80	0.884294396	0.758402852	0.890685306	0.867470239	0.866100589	0.830522324	0.813954187	0.897038611	0.887593557
booth_2	110	0.864647389	0.737585741	0.900226332	0.879324656	0.854184227	0.799960892	0.813647591	0.90230526	0.901316349
booth_3	180	0.830600312	0.700188952	0.904753756	0.836907739	0.852756623	0.801151685	0.843312503	0.899016916	0.913948478
meeting_1	210	0.814554112	0.723503635	0.886031638	0.825657718	0.822174565	0.815301842	0.800168597	0.871922117	0.888855429
meeting_2	220	0.811719495	0.713115317	0.886072635	0.817099605	0.817758927	0.820184658	0.807158072	0.867061184	0.889607494
meeting_3	240	0.799247306	0.715310619	0.874383298	0.807373563	0.812406049	0.808490548	0.800233354	0.852950258	0.880027923
meeting_4	250	0.781920327	0.710226609	0.869830816	0.802148504	0.809082253	0.806794004	0.800985712	0.854299441	0.876494749
office_1	370	0.731749054	0.678911743	0.823425193	0.759451523	0.771947619	0.778447732	0.762681199	0.820062613	0.848879194
office_2	440	0.683011943	0.623609061	0.773640501	0.703108927	0.709932035	0.729156138	0.754287925	0.771034243	0.790700449
office_3	480	0.656439105	0.587365912	0.747912266	0.669325324	0.669946009	0.689421683	0.692364471	0.735761934	0.760750571
lecture_1	700	0.728754707	0.651976944	0.840649209	0.747125397	0.792886205	0.767765779	0.801899978	0.821714938	0.859330358
lecture_2	720	0.666591082	0.603423602	0.766941068	0.686290076	0.716849377	0.717507053	0.732453516	0.759174401	0.787349193
lecture_3	790	0.650537957	0.591556585	0.751561617	0.677046222	0.690459243	0.700803302	0.704988301	0.741403842	0.762082238
lecture_4	800	0.637459188	0.579385342	0.740956208	0.665337771	0.683457717	0.680481262	0.697368467	0.737376962	0.766326057
lecture_5	810	0.628390054	0.567540791	0.727114846	0.650262432	0.663787845	0.666286281	0.669715089	0.706279586	0.738209409
lecture_6	830	0.628096751	0.573800536	0.730959875	0.655395916	0.65968295	0.673464922	0.688946224	0.704745386	0.733941737
lecture_hfrp	818	0.839198799	0.750764845	0.892655228	0.818964385	0.865067107	0.841491563	0.806736784	0.880386468	0.891498611
lecture_hhp	236	0.898674994	0.776576038	0.906111866	0.884287763	0.897150619	0.858416036	0.858276939	0.916410475	0.907644887
meeting_hfrp	337	0.853697557	0.752813161	0.894948876	0.839318333	0.866774244	0.848009935	0.812255319	0.894190738	0.894579542
meeting_hhp	254	0.898479082	0.775421384	0.905914602	0.882488333	0.894829237	0.856359288	0.856387391	0.918693011	0.908534145
office_hfrp	390	0.765606091	0.705234914	0.845145343	0.769303868	0.798447605	0.812803106	0.739801075	0.837172871	0.852297265
office_hhp	520	0.870672627	0.753485503	0.909462077	0.84474067	0.886243213	0.854327483	0.861270754	0.907988721	0.915272875

Voicing accuracy

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.9393083	0.809527809	0.913013025	0.905283868	0.935636511	0.877088199	0.856685138	0.943682112	0.914807908
booth_1	80	0.921413477	0.799802573	0.908158593	0.889176257	0.912784091	0.864872041	0.832511808	0.92109211	0.897048298
booth_2	110	0.915255177	0.793635527	0.917036158	0.902062647	0.914857216	0.844646216	0.842164382	0.925479346	0.912887609
booth_3	180	0.887220025	0.76756485	0.923866606	0.861000286	0.912517215	0.8485417	0.863121071	0.922270297	0.927838665
meeting_1	210	0.877268863	0.782036436	0.912834495	0.863569807	0.896073423	0.87029286	0.817419982	0.905788758	0.910101028
meeting_2	220	0.87265585	0.777175792	0.91375969	0.858623432	0.884500506	0.873986414	0.823110184	0.90184894	0.911589236
meeting_3	240	0.867609239	0.78110203	0.909168369	0.855196659	0.885235392	0.865415034	0.819188253	0.891508907	0.904633256
meeting_4	250	0.864181026	0.78447631	0.910078632	0.854740716	0.891409851	0.873283451	0.826388312	0.899033107	0.905207522
office_1	370	0.822231795	0.763378112	0.887646721	0.828137407	0.848938105	0.838120479	0.802550489	0.871721917	0.885483198
office_2	440	0.809721155	0.751109389	0.87898393	0.81215699	0.822953522	0.830625294	0.80368477	0.864192967	0.86582487
office_3	480	0.795460997	0.744651536	0.867760845	0.795436973	0.807832856	0.821508937	0.75673558	0.85430334	0.855281145
lecture_1	700	0.8216477	0.755312197	0.894097939	0.817984375	0.867104827	0.840625276	0.832214948	0.872511506	0.899600225
lecture_2	720	0.797381267	0.752232327	0.874572755	0.806553887	0.830249909	0.831201986	0.790802205	0.861151649	0.872937838
lecture_3	790	0.792989601	0.752838825	0.872757345	0.813025352	0.81748754	0.829935147	0.773045447	0.855068583	0.856020291
lecture_4	800	0.79057751	0.753268037	0.869107804	0.808233475	0.809774744	0.828472784	0.76228458	0.854082488	0.862941314
lecture_5	810	0.788994524	0.753120983	0.864890298	0.804962897	0.804649932	0.825655593	0.74962009	0.847440148	0.856024835
lecture_6	830	0.789994389	0.757814947	0.86318927	0.810348421	0.810044273	0.836005313	0.767459789	0.848309041	0.85176117
lecture_hfrp	818	0.883116121	0.79308549	0.906318533	0.847738247	0.910058856	0.876022425	0.818521991	0.905500212	0.90074392
lecture_hhp	236	0.932509989	0.806166418	0.916213362	0.901923125	0.933238746	0.888127458	0.868673686	0.93435834	0.913210425
meeting_hfrp	337	0.902198497	0.800051837	0.911679712	0.869230116	0.917810481	0.886878992	0.828117304	0.91999808	0.905912631
meeting_hhp	254	0.932311957	0.804926731	0.915838661	0.89975209	0.932196379	0.885105347	0.867938164	0.938544673	0.914794744
office_hfrp	390	0.837999388	0.776229291	0.890851773	0.827586769	0.868013389	0.863552033	0.775398325	0.881994792	0.877029912
office_hhp	520	0.906737523	0.788071965	0.921406564	0.866141425	0.924785664	0.884363218	0.87212648	0.928285921	0.922203166

Voicing false alarm

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.100028113	0.654046238	0.003460538	0.296280634	0.095285863	0.40890406	0.005605925	0.162899837	0.003458967
booth_1	80	0.119027113	0.690917582	0.006147571	0.329407382	0.131480452	0.447928704	0.002663117	0.19929224	0.005718047
booth_2	110	0.142990795	0.713776743	0.009824121	0.301307455	0.136192245	0.512009268	0.003953221	0.205006813	0.007990354
booth_3	180	0.210425695	0.786976889	0.011138092	0.408157867	0.145449458	0.511267391	0.006574523	0.218302911	0.008437174
meeting_1	210	0.293576625	0.763513405	0.056964897	0.471584445	0.208873073	0.452277203	0.00596444	0.292718921	0.047585691
meeting_2	220	0.301007286	0.774796827	0.065815955	0.485866586	0.227728723	0.437112963	0.004917418	0.3146194	0.06028094
meeting_3	240	0.31825207	0.76492104	0.081831378	0.499052872	0.215677431	0.455490821	0.005942284	0.33418682	0.071736934
meeting_4	250	0.323059635	0.755138053	0.08451621	0.498633297	0.221355524	0.427034002	0.007317908	0.330963224	0.078462148
office_1	370	0.442947432	0.827316833	0.17738566	0.592140968	0.341806229	0.57265665	0.015363967	0.443179113	0.163557275
office_2	440	0.452740718	0.859369246	0.216717135	0.624252677	0.383122113	0.596493347	0.046957226	0.434458623	0.234416108
office_3	480	0.463335297	0.876239862	0.237468317	0.658859512	0.384335168	0.623771645	0.041228729	0.457873289	0.23532029
lecture_1	700	0.436586272	0.846516378	0.184307927	0.625006449	0.295396114	0.555202036	0.012290994	0.432171051	0.159870108
lecture_2	720	0.470815135	0.855693511	0.240727405	0.639407172	0.369197998	0.6014701	0.050943624	0.460743756	0.24755162
lecture_3	790	0.4783158	0.854913935	0.234191782	0.631815851	0.363813825	0.589420234	0.049886931	0.454962279	0.235270326
lecture_4	800	0.48291568	0.853659247	0.253067258	0.646574524	0.410193702	0.597145511	0.047039217	0.478790153	0.252587101
lecture_5	810	0.482874455	0.852580183	0.258285513	0.650240304	0.390940971	0.602034151	0.073251916	0.486553099	0.269739078
lecture_6	830	0.4918697	0.840348002	0.254422167	0.639022346	0.383454459	0.572029336	0.054256084	0.490160037	0.258139933
lecture_hfrp	818	0.278519577	0.713678228	0.035639355	0.532957403	0.181350843	0.408362629	0.003637233	0.311745927	0.025593587
lecture_hhp	236	0.134439767	0.674332402	0.005493189	0.327735905	0.125514571	0.379972274	0.004857398	0.197264214	0.00420969
meeting_hfrp	337	0.224083615	0.690531816	0.033186387	0.448878312	0.147776538	0.372803726	0.005336834	0.263439705	0.022632456
meeting_hhp	254	0.135242374	0.678500164	0.005737206	0.334885911	0.129288946	0.388659295	0.004322263	0.191662495	0.006441392
office_hfrp	390	0.393219972	0.778380548	0.109248595	0.597544816	0.254214869	0.457701565	0.007582783	0.394396932	0.09350951
office_hhp	520	0.219469382	0.735306568	0.01734692	0.464309652	0.157718808	0.392993118	0.004103794	0.2419343	0.009156791

Voicing detection

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.952218869	0.99834949	0.877206184	0.983764904	0.945221656	0.993422645	0.799379415	0.981320467	0.881185376
booth_1	80	0.935024116	0.998893466	0.871451448	0.973349709	0.926331723	0.989922333	0.768870735	0.962661494	0.857941369
booth_2	110	0.935051228	0.998756687	0.88500833	0.979299795	0.930932451	0.993565379	0.781586283	0.972581301	0.881071892
booth_3	180	0.930011695	0.998597197	0.895627602	0.969414964	0.931126284	0.996095496	0.81026567	0.970823794	0.902185633
meeting_1	210	0.939791958	0.999461745	0.898301279	0.989088742	0.931302749	0.995331626	0.745701378	0.9777956	0.892482131
meeting_2	220	0.936100338	0.999422708	0.903286373	0.98879118	0.922070685	0.993105621	0.75477012	0.978245977	0.897307874
meeting_3	240	0.935420905	0.999346123	0.902640507	0.988807183	0.91922015	0.99288699	0.749666584	0.973144782	0.895276451
meeting_4	250	0.93337471	0.999228255	0.905286482	0.987672907	0.929976356	0.993211372	0.760016195	0.98097239	0.898374488
office_1	370	0.920866693	0.999881864	0.909325182	0.988666668	0.913800495	0.995720025	0.730348793	0.986466042	0.904536993
office_2	440	0.906124252	0.999755517	0.912133292	0.98018421	0.89113238	0.995213582	0.741047149	0.973694311	0.902724925
office_3	480	0.889388764	0.999717095	0.904248085	0.97238807	0.869454222	0.996677279	0.67829295	0.969450509	0.88738873
lecture_1	700	0.918021307	0.999636871	0.921997411	0.988279754	0.921491971	0.993691231	0.771199381	0.983099103	0.921018241
lecture_2	720	0.896709432	0.999517629	0.915845853	0.9785501	0.900597679	0.997071567	0.728729991	0.979553106	0.916697057
lecture_3	790	0.89295169	0.999426077	0.910519066	0.983895489	0.878217704	0.99226865	0.69849305	0.968194943	0.88816929
lecture_4	800	0.891640936	0.999238353	0.912664809	0.983154042	0.881830643	0.991237533	0.687435524	0.975098541	0.902654568
lecture_5	810	0.890041768	0.99900244	0.90861515	0.979786919	0.870775525	0.990781058	0.678370759	0.968063823	0.900731603
lecture_6	830	0.895135265	0.999062634	0.904758426	0.983013126	0.87436037	0.994217256	0.695668593	0.971494698	0.891300717
lecture_hfrp	818	0.943840102	0.998232435	0.880584553	0.991570136	0.94025705	0.991898682	0.748647757	0.982773963	0.870133601
lecture_hhp	236	0.956257672	0.999126701	0.882169329	0.988435464	0.952902921	0.995462413	0.816868329	0.98216029	0.880829327
meeting_hfrp	337	0.947858964	0.998380837	0.887164971	0.989802892	0.939572244	0.991794127	0.760571977	0.983640323	0.875706291
meeting_hhp	254	0.956110877	0.999119377	0.881764866	0.988284837	0.952747302	0.995030105	0.815621519	0.985007213	0.883274676
office_hfrp	390	0.924124451	0.998902126	0.888286418	0.989502036	0.908738706	0.988055526	0.688433533	0.98209501	0.864011605
office_hhp	520	0.954191685	0.999366254	0.894239703	0.99084467	0.95273268	0.994525359	0.821191803	0.988138108	0.894751902

Raw pitch accuracy

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.931407144	0.946640946	0.874813882	0.959897325	0.928746408	0.951494659	0.804887625	0.955573669	0.955573669
booth_1	80	0.906010624	0.938646251	0.854951401	0.942002584	0.885202489	0.941338948	0.762969939	0.928899823	0.928899823
booth_2	110	0.885301712	0.917196182	0.870231751	0.946584103	0.869419305	0.930151084	0.764675474	0.939632644	0.939632644
booth_3	180	0.871871732	0.901781254	0.877547625	0.934774108	0.871172032	0.929224669	0.80531877	0.938139343	0.938139343
meeting_1	210	0.871604272	0.914929208	0.86624747	0.93479008	0.856291072	0.919437706	0.748037467	0.930157987	0.930157987
meeting_2	220	0.871925287	0.907096491	0.869011036	0.929342474	0.859243796	0.916957392	0.755457366	0.928933532	0.928933532
meeting_3	240	0.860962882	0.904567411	0.860780498	0.920385002	0.845449775	0.913415634	0.756842592	0.918489487	0.918489487
meeting_4	250	0.839466836	0.892657446	0.855635349	0.912605279	0.839460312	0.900427623	0.75467668	0.918385181	0.918385181
office_1	370	0.820970155	0.878660704	0.824170261	0.890869036	0.846466058	0.912919283	0.6937318	0.913343758	0.913343758
office_2	440	0.756545994	0.817359076	0.771965857	0.824414855	0.775534056	0.853098096	0.707638588	0.842628112	0.842628112
office_3	480	0.726212033	0.774516889	0.745381191	0.792101233	0.723957108	0.811123754	0.633805155	0.802785022	0.802785022
lecture_1	700	0.81423438	0.851197876	0.851953194	0.88701515	0.849779059	0.891252861	0.751466663	0.911493261	0.911493261
lecture_2	720	0.744727951	0.786869468	0.76842342	0.807414022	0.778572336	0.839286904	0.670221079	0.837714192	0.837714192
lecture_3	790	0.724826388	0.768495842	0.744961923	0.78998984	0.750036242	0.81119036	0.626914444	0.809066291	0.809066291
lecture_4	800	0.706701817	0.750588558	0.736869362	0.77949836	0.748328901	0.784637302	0.621873626	0.812341277	0.812341277
lecture_5	810	0.692932137	0.733524916	0.718803132	0.759197315	0.718492074	0.767132393	0.59220831	0.770654615	0.770654615
lecture_6	830	0.69369118	0.735881051	0.726511632	0.761781129	0.703268461	0.76688544	0.624968882	0.769758078	0.769758078
lecture_hfrp	818	0.902387043	0.937026975	0.870068016	0.9502053	0.903005579	0.943273614	0.757669668	0.947054349	0.947054349
lecture_hhp	236	0.924081496	0.956312715	0.875772689	0.962990346	0.919715043	0.953528423	0.819138278	0.956612393	0.956612393
meeting_hfrp	337	0.89722798	0.930133842	0.871387272	0.947068635	0.892149403	0.937501799	0.763281294	0.947391189	0.947391189
meeting_hhp	254	0.923957236	0.956426563	0.875308272	0.963405039	0.917423849	0.9541403	0.814736432	0.95679753	0.95679753
office_hfrp	390	0.849488875	0.897095282	0.835153133	0.906244226	0.849440332	0.916801474	0.683988881	0.918509775	0.918509775
office_hhp	520	0.91978378	0.949476397	0.884319789	0.960065053	0.916816149	0.95166497	0.822282595	0.959141539	0.959141539

Voiced pitch accuracy

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.911834824	0.946640946	0.866175817	0.959897325	0.905548301	0.951494659	0.786628784	0.955573669	0.873487836
booth_1	80	0.882064577	0.938646251	0.846263858	0.942002584	0.861505443	0.941338948	0.743027881	0.928899823	0.844838432
booth_2	110	0.863139918	0.917196182	0.860545864	0.946584103	0.84671804	0.930151084	0.741240284	0.939632644	0.864482066
booth_3	180	0.849428876	0.901781254	0.868250736	0.934774108	0.84804101	0.929224669	0.782612236	0.938139343	0.882766443
meeting_1	210	0.850585775	0.914929208	0.859851885	0.93479008	0.829492683	0.919437706	0.721428253	0.930157987	0.862760147
meeting_2	220	0.849247191	0.907096491	0.863550593	0.929342474	0.829460761	0.916957392	0.731966117	0.928933532	0.866367003
meeting_3	240	0.838228546	0.904567411	0.852575537	0.920385002	0.818407536	0.913415634	0.723212163	0.918489487	0.8597059
meeting_4	250	0.816307923	0.892657446	0.847628302	0.912605279	0.816401762	0.900427623	0.724212887	0.918385181	0.857876443
office_1	370	0.791828364	0.878660704	0.817736043	0.890869036	0.807088487	0.912919283	0.673870748	0.913343758	0.852569183
office_2	440	0.725386839	0.817359076	0.761864203	0.824414855	0.73250817	0.853098096	0.672336251	0.842628112	0.796520251
office_3	480	0.690651926	0.774516889	0.732686464	0.792101233	0.676446101	0.811123754	0.587185276	0.802785022	0.754225876
lecture_1	700	0.785641199	0.851197876	0.845464696	0.88701515	0.819412439	0.891252861	0.730559406	0.911493261	0.864009635
lecture_2	720	0.710313245	0.786869468	0.762630647	0.807414022	0.741632654	0.839286904	0.64568047	0.837714192	0.797399404
lecture_3	790	0.68968179	0.768495842	0.73796899	0.78998984	0.700642199	0.81119036	0.603504991	0.809066291	0.756227201
lecture_4	800	0.673358473	0.750588558	0.730015042	0.77949836	0.705149023	0.784637302	0.596536325	0.812341277	0.767903941
lecture_5	810	0.660869923	0.733524916	0.711853516	0.759197315	0.673681316	0.767132393	0.565241572	0.770654615	0.735590915
lecture_6	830	0.663949585	0.735881051	0.716192918	0.761781129	0.66332447	0.76688544	0.585358194	0.769758078	0.725218203
lecture_hfrp	818	0.881392538	0.937026975	0.860868868	0.9502053	0.878731228	0.943273614	0.732328891	0.947054349	0.857039528
lecture_hhp	236	0.908453763	0.956312715	0.867549075	0.962990346	0.903825213	0.953528423	0.802329541	0.956612393	0.872902231
meeting_hfrp	337	0.87914499	0.930133842	0.8629908	0.947068635	0.869622276	0.937501799	0.738343624	0.947391189	0.860032338
meeting_hhp	254	0.908305877	0.956426563	0.867465636	0.963405039	0.901844645	0.9541403	0.799467742	0.95679753	0.874413167
office_hfrp	390	0.820761331	0.897095282	0.822508496	0.906244226	0.811861681	0.916801474	0.638661905	0.918509775	0.828809167
office_hhp	520	0.903229375	0.949476397	0.877005624	0.960065053	0.90038824	0.95166497	0.805893061	0.959141539	0.88485683

F-measure

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.939125742	0.852140022	0.946130905	0.931791842	0.93669835	0.901235452	0.912291478	0.951875501	0.94928794
booth_1	80	0.919830599	0.839706367	0.932644721	0.917299629	0.904890072	0.88891117	0.882490431	0.933336409	0.934680671
booth_2	110	0.902267075	0.819141655	0.938674108	0.923944601	0.891761242	0.864554167	0.877380267	0.936902693	0.941944902
booth_3	180	0.878943223	0.788806976	0.940089346	0.897236211	0.89137612	0.864649476	0.904004408	0.935017213	0.947955059
meeting_1	210	0.867005034	0.809430162	0.92588994	0.885339568	0.866887545	0.870612159	0.877722113	0.914544655	0.930132915
meeting_2	220	0.865964509	0.799869687	0.925375324	0.87831236	0.867207777	0.87468053	0.882157915	0.911281596	0.930287494
meeting_3	240	0.855026411	0.800572036	0.915504804	0.869513922	0.861012193	0.865688048	0.875698196	0.901218173	0.92336564
meeting_4	250	0.838002966	0.793609439	0.910346378	0.864019916	0.855113929	0.859851302	0.871981557	0.899191457	0.919358119
office_1	370	0.801721271	0.767558266	0.871625168	0.829707119	0.834106564	0.845756801	0.841713859	0.875659302	0.899562739
office_2	440	0.753141501	0.7095845	0.822698273	0.774377948	0.777871198	0.794605279	0.833265108	0.826026302	0.846322755
office_3	480	0.729323484	0.670382693	0.799974482	0.743451215	0.739942251	0.755259699	0.784562471	0.792613886	0.818918927
lecture_1	700	0.798737213	0.740233852	0.88702196	0.820675886	0.848572029	0.833220277	0.872382941	0.876695032	0.90436956
lecture_2	720	0.740036427	0.686287449	0.817471806	0.758253388	0.782478555	0.782248838	0.814319908	0.815001067	0.840370817
lecture_3	790	0.72380843	0.672610474	0.801699228	0.745685897	0.758855046	0.764050268	0.790917175	0.798423518	0.820199544
lecture_4	800	0.710158602	0.658624066	0.791949017	0.73457275	0.754227775	0.742867825	0.78741245	0.794677754	0.821985641
lecture_5	810	0.700680195	0.645228537	0.778904059	0.719134456	0.734790704	0.728410524	0.763430005	0.763696622	0.794398713
lecture_6	830	0.699943126	0.650548779	0.783625723	0.72280188	0.728040309	0.732259421	0.777550662	0.761675917	0.79136473
lecture_hfrp	818	0.89056336	0.834959765	0.935771026	0.885537613	0.905382506	0.895129658	0.883209196	0.923583025	0.937325084
lecture_hhp	236	0.929747922	0.857271453	0.945381528	0.929244688	0.927862988	0.907671869	0.917174585	0.947209912	0.948518786
meeting_hfrp	337	0.897073903	0.833874441	0.935756684	0.8972913	0.903575177	0.897479212	0.88420838	0.931166861	0.938076407
meeting_hhp	254	0.929647638	0.856563172	0.945358497	0.928408062	0.925975771	0.907005431	0.915690033	0.947587257	0.94867461
office_hfrp	390	0.831794696	0.791774518	0.892502856	0.840664869	0.853952539	0.871090912	0.825350191	0.889061868	0.907033014
office_hhp	520	0.912794091	0.840215866	0.946345065	0.904462393	0.920607208	0.90518504	0.918704655	0.941462261	0.9520894

Precision

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.941764283	0.774580449	0.990510253	0.898432415	0.943107858	0.851338561	0.987794899	0.937370646	0.99283092
booth_1	80	0.925696084	0.759023656	0.978272999	0.883370483	0.913115591	0.836744247	0.97547493	0.920209572	0.987450566
booth_2	110	0.904633372	0.738265421	0.978664049	0.891852462	0.897607982	0.803813233	0.958686497	0.920038966	0.985081542
booth_3	180	0.872106292	0.70092231	0.975884403	0.854316082	0.896092005	0.803397512	0.975238292	0.917403664	0.982461967
meeting_1	210	0.849527874	0.723793451	0.95405585	0.831769582	0.864027291	0.818026203	0.977329948	0.885556868	0.963344437
meeting_2	220	0.848683884	0.713414387	0.950416789	0.823265992	0.864988129	0.824337496	0.979111769	0.880427145	0.959521085
meeting_3	240	0.835971936	0.715648743	0.938547685	0.813522942	0.860755773	0.812490067	0.97451787	0.868782972	0.951850703
meeting_4	250	0.819101623	0.710627839	0.931768022	0.808789143	0.85107846	0.810742507	0.96564158	0.865932343	0.946405456
office_1	370	0.77400941	0.678966028	0.879096313	0.765262869	0.822490597	0.780997109	0.942804248	0.827796137	0.912558037
office_2	440	0.729879982	0.623717456	0.82418271	0.712398929	0.769401104	0.731718662	0.926621191	0.7851028	0.850349775
office_3	480	0.709819732	0.587487841	0.801475885	0.681572525	0.737549215	0.691119938	0.899512135	0.752147377	0.827731785
lecture_1	700	0.772284948	0.652147437	0.889553029	0.75308851	0.839418353	0.771331175	0.959170855	0.83137806	0.911775344
lecture_2	720	0.717796046	0.603629534	0.8149182	0.696178342	0.771737036	0.718956435	0.911050098	0.770466568	0.83859724
lecture_3	790	0.702248035	0.591787349	0.80150326	0.684385836	0.756099453	0.704814037	0.894580099	0.758640707	0.829295898
lecture_4	800	0.688583369	0.579686545	0.789071502	0.672859159	0.746029553	0.685004086	0.897008752	0.750442867	0.823598848
lecture_5	810	0.679981794	0.567934756	0.7768227	0.659110264	0.731243462	0.670898004	0.872236785	0.722297402	0.794580088
lecture_6	830	0.676847243	0.57417364	0.782938239	0.662947353	0.724301054	0.676239295	0.880372061	0.719215556	0.79579004
lecture_hfrp	818	0.872970365	0.751740896	0.974256802	0.823682627	0.903352186	0.846344797	0.981852304	0.891334533	0.98287236
lecture_hhp	236	0.926390075	0.777079281	0.987384336	0.891269997	0.927662376	0.861058512	0.986351649	0.928279324	0.992717614
meeting_hfrp	337	0.885254411	0.753698298	0.971975531	0.845134867	0.905419367	0.853105772	0.977039236	0.904468023	0.981194483
meeting_hhp	254	0.926288233	0.775928532	0.987535313	0.889556712	0.925385452	0.859286243	0.98512609	0.928653233	0.991509811
office_hfrp	390	0.807649714	0.705804005	0.916226229	0.774754066	0.853533493	0.820080485	0.94442469	0.848168059	0.944979888
office_hhp	520	0.898904522	0.75384077	0.982183153	0.850016194	0.916653033	0.857720392	0.986146569	0.915656621	0.990092918

Recall

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.936943412	0.953633674	0.90630817	0.96947688	0.930685051	0.960915054	0.849015092	0.967256744	0.910132145
booth_1	80	0.915128519	0.946092457	0.892283449	0.95549911	0.897801993	0.9521652	0.814742759	0.947205524	0.889105522
booth_2	110	0.900601403	0.926396654	0.90272884	0.959444652	0.88633778	0.941369203	0.814615793	0.95487139	0.903343718
booth_3	180	0.888415574	0.910124132	0.907579107	0.94825811	0.887076751	0.941063684	0.844840252	0.953809217	0.91625406
meeting_1	210	0.88625943	0.923651827	0.900354992	0.947708794	0.870592793	0.933349253	0.801431286	0.946060324	0.900336642
meeting_2	220	0.885061084	0.916203107	0.902658733	0.94288624	0.87036364	0.932980901	0.808394442	0.945408368	0.903712525
meeting_3	240	0.876088287	0.91408982	0.894565933	0.935410341	0.862155604	0.92903382	0.801640785	0.936816418	0.898015444
meeting_4	250	0.859050938	0.903868212	0.890926744	0.929060496	0.859856824	0.918653716	0.802498378	0.935870695	0.89565979
office_1	370	0.834064366	0.888334405	0.865765335	0.908300347	0.84855169	0.925629294	0.766168053	0.930500497	0.889134123
office_2	440	0.780363074	0.829448563	0.822623148	0.850756108	0.7891473	0.872690208	0.763592131	0.872629574	0.844198051
office_3	480	0.752566219	0.787307231	0.799919149	0.820933959	0.744009368	0.835735843	0.700603207	0.839329089	0.81208893
lecture_1	700	0.82942644	0.862146184	0.885959352	0.904089739	0.859387519	0.908437492	0.804898121	0.928559971	0.898143548
lecture_2	720	0.766811041	0.801258947	0.821610045	0.8351228	0.795658764	0.860298155	0.742401214	0.866028012	0.843582363
lecture_3	790	0.749896498	0.784658222	0.803278343	0.82136567	0.764233119	0.837420147	0.714130464	0.844056159	0.813568595
lecture_4	800	0.7359151	0.768089636	0.796325122	0.811272151	0.764739431	0.814304551	0.705938096	0.845578497	0.821989335
lecture_5	810	0.725712172	0.752270705	0.782472303	0.793714999	0.741266613	0.799474325	0.683037986	0.81115005	0.796019192
lecture_6	830	0.727428555	0.755499441	0.785800415	0.796814432	0.733791818	0.801107098	0.699209987	0.810686224	0.788684393
lecture_hfrp	818	0.910175613	0.94450093	0.901668733	0.959358678	0.908146269	0.954075659	0.807614138	0.95905152	0.897290469
lecture_hhp	236	0.933505107	0.962019389	0.907472966	0.971729724	0.928430672	0.96305962	0.859378546	0.967366594	0.908660659
meeting_hfrp	337	0.909871559	0.93891116	0.903270928	0.957753505	0.902373621	0.949576319	0.813356063	0.959906025	0.899766992
meeting_hhp	254	0.933401107	0.96203778	0.907303049	0.971952333	0.926918207	0.963676269	0.857415066	0.967650505	0.909961651
office_hfrp	390	0.859535311	0.907058465	0.87205047	0.921224675	0.855926914	0.932095567	0.741437222	0.935640921	0.874509694
office_hhp	520	0.927917022	0.955156918	0.913709156	0.968109343	0.925031771	0.961645513	0.862224137	0.969142465	0.917358139

Overall chroma accuracy

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.927714548	0.788869605	0.906587496	0.893395498	0.924909257	0.86253346	0.850135789	0.933338067	0.907836509
booth_1	80	0.909066693	0.775627323	0.900522262	0.873900914	0.89958591	0.845930094	0.825776343	0.908604259	0.889305506
booth_2	110	0.899225517	0.763817009	0.908624168	0.889090507	0.896932016	0.821020085	0.833629339	0.912137711	0.904379963
booth_3	180	0.86950823	0.727533511	0.915279646	0.84563488	0.896835084	0.820874693	0.854855678	0.909714889	0.918695748
meeting_1	210	0.858930313	0.753636327	0.899594379	0.843812863	0.878811817	0.844911293	0.809042408	0.890762294	0.900056349
meeting_2	220	0.855060537	0.745916343	0.90251534	0.83721287	0.867802971	0.848930276	0.815906359	0.886756458	0.901018127
meeting_3	240	0.848814728	0.748816982	0.895538172	0.832223799	0.870186333	0.841715312	0.813942059	0.871573712	0.890791008
meeting_4	250	0.840985339	0.748635398	0.89261252	0.828903388	0.871449879	0.842713522	0.812695307	0.878688594	0.890501183
office_1	370	0.796628807	0.719967419	0.864439891	0.797750032	0.826094975	0.809044375	0.790835778	0.849457418	0.869075006
office_2	440	0.775187207	0.688849523	0.84165657	0.767934929	0.7933811	0.78514223	0.788873123	0.828081526	0.837901111
office_3	480	0.755293116	0.665331233	0.823846725	0.741802008	0.778066779	0.762516238	0.741231634	0.81395701	0.825506467
lecture_1	700	0.796156429	0.703947015	0.87223517	0.785173579	0.847755317	0.810640222	0.820810975	0.851753253	0.884157434
lecture_2	720	0.760838536	0.678352495	0.834609227	0.756357412	0.799490084	0.780404715	0.772522009	0.825430869	0.843184404
lecture_3	790	0.753352745	0.676415824	0.832360385	0.761005828	0.785614414	0.778561945	0.751868556	0.815380718	0.823918998
lecture_4	800	0.746094456	0.672108128	0.825465595	0.753218495	0.776741199	0.767812421	0.743606767	0.811881071	0.830271172
lecture_5	810	0.744233199	0.668469887	0.819316517	0.747385269	0.770872201	0.761259613	0.72697517	0.802331534	0.819591591
lecture_6	830	0.745963605	0.67236438	0.821462369	0.75199608	0.778850198	0.770491562	0.749262523	0.803134437	0.817040594
lecture_hfrp	818	0.868909097	0.771415517	0.898279752	0.830671898	0.89670074	0.859164768	0.809662084	0.889695815	0.891477818
lecture_hhp	236	0.919856623	0.790362232	0.90942545	0.889802281	0.921539746	0.872657178	0.861429327	0.923095958	0.905604067
meeting_hfrp	337	0.886376221	0.776335033	0.90280787	0.852628398	0.903551056	0.868806124	0.818318413	0.90635094	0.897082624
meeting_hhp	254	0.919794249	0.789414849	0.909288242	0.888007943	0.921283822	0.869669196	0.861081614	0.925445615	0.906796018
office_hfrp	390	0.81621019	0.739185694	0.872700151	0.798685599	0.849661736	0.838015905	0.75643327	0.859662359	0.863006374
office_hhp	520	0.893902193	0.770058033	0.914429649	0.852080455	0.913073193	0.871077139	0.86453645	0.91503912	0.913850942

Raw chroma accuracy

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.969673454	0.969019864	0.897903243	0.968728003	0.968819543	0.973680761	0.855699595	0.968179727	0.968179727
booth_1	80	0.961385043	0.964232304	0.888004706	0.955737218	0.952759476	0.964888665	0.822715012	0.950771192	0.950771192
booth_2	110	0.955205029	0.956019877	0.900886249	0.963017546	0.949604142	0.960666278	0.843278528	0.955746476	0.955746476
booth_3	180	0.949896632	0.941639153	0.911834334	0.95096274	0.949565598	0.957531344	0.863708166	0.955489257	0.955489257
meeting_1	210	0.955041088	0.958891446	0.903432447	0.961870413	0.955117167	0.96099367	0.823516351	0.958979863	0.958979863
meeting_2	220	0.954834208	0.954936414	0.907457747	0.959583312	0.952992522	0.958915616	0.821843511	0.958233203	0.958233203
meeting_3	240	0.952504632	0.95345806	0.906191322	0.957692648	0.948860507	0.96067753	0.819562337	0.949426506	0.949426506
meeting_4	250	0.945820841	0.948174958	0.904021859	0.952802126	0.949717931	0.951551852	0.824083998	0.954307964	0.954307964
office_1	370	0.939794434	0.938101036	0.897881041	0.946667198	0.946031673	0.955727313	0.79439085	0.95626983	0.95626983
office_2	440	0.919102866	0.911095673	0.888792958	0.919503092	0.92752383	0.932588771	0.816046249	0.924838935	0.924838935
office_3	480	0.905347952	0.886622613	0.875629582	0.899189882	0.917571242	0.915662124	0.779150251	0.915713665	0.915713665
lecture_1	700	0.938794825	0.926452368	0.911740157	0.943049985	0.949352864	0.952508434	0.825604036	0.955279982	0.955279982
lecture_2	720	0.913757014	0.89452012	0.87704334	0.909574564	0.925911276	0.927424196	0.771968356	0.930924505	0.930924505
lecture_3	790	0.907539762	0.890658997	0.878309509	0.911674526	0.921162892	0.922247074	0.775989682	0.915903451	0.915903451
lecture_4	800	0.89929663	0.883648374	0.872770103	0.906708418	0.913765225	0.909053643	0.7438827	0.920077486	0.920077486
lecture_5	810	0.89656037	0.878559662	0.8703649	0.899717264	0.910361778	0.902976272	0.760464071	0.908736287	0.908736287
lecture_6	830	0.897662702	0.877513104	0.872728115	0.902273583	0.912635541	0.905744469	0.769409114	0.912392816	0.912392816
lecture_hfrp	818	0.962985383	0.967533291	0.901092949	0.968125767	0.960994364	0.969426593	0.822937556	0.962865056	0.962865056
lecture_hhp	236	0.967841977	0.976637792	0.901763455	0.972662549	0.96800179	0.974484644	0.867010203	0.967530493	0.967530493
meeting_hfrp	337	0.961541585	0.964691379	0.907404382	0.967171838	0.961669445	0.968654231	0.833270186	0.966369622	0.966369622
meeting_hhp	254	0.968109204	0.977048167	0.900334214	0.973117522	0.967768776	0.973723686	0.862355463	0.967762358	0.967762358
office_hfrp	390	0.946701153	0.946314761	0.893390734	0.949546164	0.946202585	0.953401206	0.770748751	0.95227689	0.95227689
office_hhp	520	0.967674768	0.973707142	0.910144062	0.971883008	0.96647935	0.976517741	0.86586556	0.970636697	0.970636697

Voiced chroma accuracy

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.935760189	0.968758965	0.868025045	0.966754144	0.930112993	0.972588618	0.789949518	0.966656271	0.871166822
booth_1	80	0.917548823	0.963996098	0.860579145	0.951533452	0.90765206	0.963017728	0.759341814	0.944803885	0.847058298
booth_2	110	0.912311338	0.955851779	0.873061629	0.960692076	0.905788429	0.959690577	0.769431451	0.95328458	0.868677125
booth_3	180	0.90481477	0.941458619	0.883457699	0.947499475	0.908925541	0.956968439	0.798490976	0.95296685	0.889104301
meeting_1	210	0.913832954	0.958816842	0.879411663	0.960956347	0.90726775	0.960224367	0.733779155	0.956434586	0.878140919
meeting_2	220	0.911214972	0.95485568	0.887223516	0.958314029	0.898904717	0.957837623	0.744502117	0.956808178	0.88217425
meeting_3	240	0.908828229	0.953359593	0.883213813	0.956113545	0.898003808	0.959666798	0.741971141	0.944958847	0.875303094
meeting_4	250	0.900493746	0.948054497	0.880428384	0.95098117	0.90217524	0.95054112	0.740711196	0.952493792	0.877513695
office_1	370	0.884572934	0.938088447	0.876329102	0.945543339	0.881733356	0.955239462	0.714024062	0.954903051	0.881047293
office_2	440	0.857268125	0.911068417	0.85920539	0.9173673	0.849574746	0.931949281	0.72043313	0.922827374	0.863107697
office_3	480	0.832361172	0.886590305	0.841768778	0.896015768	0.828088574	0.914651391	0.656398017	0.913295869	0.845635035
lecture_1	700	0.881955358	0.92641437	0.890743397	0.941613467	0.894416806	0.95194553	0.754595477	0.95370962	0.898910337
lecture_2	720	0.844922633	0.894444908	0.859149621	0.907294178	0.857953842	0.926619862	0.702386441	0.92945243	0.87486922
lecture_3	790	0.836766403	0.890616333	0.853178492	0.910192982	0.833696635	0.920657966	0.668696863	0.912696628	0.843098982
lecture_4	800	0.82859054	0.883561233	0.850682544	0.904991116	0.836081224	0.907107192	0.661338133	0.916405738	0.857203381
lecture_5	810	0.826423351	0.878438806	0.844010984	0.898033539	0.823809407	0.901132255	0.646331327	0.905050777	0.849649019
lecture_6	830	0.83264351	0.87740165	0.845422133	0.899973334	0.830828444	0.903481221	0.670003364	0.908230326	0.842388211
lecture_hfrp	818	0.923770395	0.96733034	0.869066401	0.967196757	0.921370274	0.968010575	0.736251357	0.960193208	0.856813444
lecture_hhp	236	0.938364205	0.976506482	0.872514198	0.971136214	0.936388933	0.973508943	0.806590724	0.966087103	0.869913066
meeting_hfrp	337	0.925514915	0.964517578	0.874504229	0.966138066	0.919375683	0.966666393	0.74673547	0.964169511	0.863222796
meeting_hhp	254	0.938416099	0.976912198	0.872444742	0.971546393	0.93731744	0.972954383	0.805925591	0.966340305	0.87181729
office_hfrp	390	0.893245754	0.946161025	0.862283916	0.948501412	0.882932217	0.951728657	0.662006231	0.950086831	0.843720594
office_hhp	520	0.936036579	0.973617622	0.884324079	0.970719083	0.935964604	0.975542039	0.810402164	0.969042931	0.882661567

Chroma f-measure

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.956546121	0.868981923	0.947415278	0.936768639	0.955442065	0.917465991	0.914818306	0.960121656	0.947836648
booth_1	80	0.945757256	0.85901555	0.943142931	0.924174952	0.940029709	0.905527978	0.895757137	0.945540273	0.936620969
booth_2	110	0.938554944	0.84871293	0.947478061	0.934300592	0.936509994	0.887231758	0.900812297	0.947211886	0.94513565
booth_3	180	0.92020293	0.819834142	0.951111438	0.90664925	0.937569796	0.885877379	0.916325031	0.946200656	0.952852887
meeting_1	210	0.914500778	0.843423907	0.940177987	0.904904173	0.926714343	0.902290592	0.88738074	0.934391908	0.941897549
meeting_2	220	0.912499831	0.836922079	0.942713926	0.900077317	0.920500348	0.905502206	0.891560989	0.932138329	0.942195988
meeting_3	240	0.908278746	0.838408389	0.937734782	0.896401562	0.922588735	0.901350786	0.890846588	0.921049697	0.934631845
meeting_4	250	0.901578192	0.836735607	0.934268457	0.8929612	0.921220814	0.898345214	0.884743122	0.924909037	0.934086751
office_1	370	0.873142093	0.814240962	0.915369	0.871652782	0.892892988	0.879193143	0.873701716	0.907208879	0.921176346
office_2	440	0.855294676	0.784040987	0.895286785	0.845980463	0.869424634	0.856332183	0.871247724	0.887548666	0.897279039
office_3	480	0.839830289	0.759707	0.881542055	0.824173035	0.860063453	0.836058812	0.840573433	0.87719046	0.889162894
lecture_1	700	0.872975218	0.799479956	0.920499697	0.862578947	0.907562859	0.879960058	0.893076156	0.908994185	0.930529778
lecture_2	720	0.845132748	0.771721501	0.889830917	0.83574846	0.872763365	0.851208133	0.858929483	0.886356795	0.900078212
lecture_3	790	0.838688591	0.769366282	0.888214183	0.83827679	0.863696339	0.849488832	0.844320449	0.878441942	0.88698096
lecture_4	800	0.831770193	0.764234017	0.882602104	0.831707633	0.857670788	0.838825011	0.840659974	0.875222365	0.890971327
lecture_5	810	0.830444736	0.760206057	0.877966508	0.826696271	0.853598998	0.83288638	0.828725721	0.867861112	0.882504057
lecture_6	830	0.831887523	0.762645995	0.881021016	0.829493572	0.860035266	0.838367752	0.84606831	0.868406843	0.881564301
lecture_hfrp	818	0.92226552	0.858117589	0.941777271	0.898246358	0.938690934	0.914133799	0.88667327	0.933473501	0.937246562
lecture_hhp	236	0.951752533	0.872627603	0.948850084	0.935116396	0.953083959	0.922922104	0.920584573	0.954166751	0.946396307
meeting_hfrp	337	0.931558977	0.860124734	0.944026279	0.91159015	0.942085278	0.919709647	0.890803835	0.943891639	0.940692062
meeting_hhp	254	0.95178512	0.872163549	0.948889918	0.934292364	0.953374541	0.921266132	0.920752069	0.954568187	0.946861568
office_hfrp	390	0.887140648	0.830063333	0.922120687	0.872895746	0.909016228	0.89808333	0.84424195	0.913037683	0.918586757
office_hhp	520	0.937231153	0.858802128	0.95152257	0.912399288	0.948536492	0.923084377	0.922231797	0.948833847	0.95060434

Chroma precision

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.959305594	0.789807654	0.991857114	0.903231064	0.961996787	0.86659077	0.990512912	0.945456425	0.991312008
booth_1	80	0.951925464	0.776259529	0.989528411	0.889970837	0.948711542	0.852229075	0.990737224	0.932189489	0.989696743
booth_2	110	0.941198118	0.764518842	0.987894797	0.901819577	0.942773947	0.824970131	0.988542325	0.930074025	0.988392742
booth_3	180	0.913345619	0.728294776	0.987440075	0.863292974	0.942529313	0.823196044	0.988472659	0.92829947	0.987501307
meeting_1	210	0.896222524	0.753938527	0.96884478	0.85009571	0.923848352	0.847766511	0.988021709	0.904641824	0.975633945
meeting_2	220	0.894471414	0.746228566	0.968341699	0.843590477	0.918358814	0.853189804	0.989305146	0.90031555	0.971703669
meeting_3	240	0.888263741	0.749172122	0.961321721	0.838606728	0.922512097	0.845878065	0.991576552	0.887747904	0.96335362
meeting_4	250	0.881505315	0.7490586	0.956266858	0.835824789	0.917133977	0.846823078	0.979801143	0.890579187	0.961636933
office_1	370	0.843178956	0.720024777	0.923487681	0.803938649	0.880566881	0.811670001	0.980319134	0.857480703	0.934607124
office_2	440	0.829097881	0.688968886	0.896954471	0.778223728	0.859791679	0.787926038	0.968255555	0.843431226	0.901722687
office_3	480	0.817749725	0.665467346	0.88323963	0.755620959	0.857656407	0.76436416	0.965118254	0.832098734	0.898905175
lecture_1	700	0.844282622	0.704132106	0.92319523	0.791492697	0.898048158	0.814432738	0.982080029	0.861752571	0.938012313
lecture_2	720	0.819962726	0.67858477	0.887124264	0.767383261	0.861057326	0.78198682	0.960946467	0.837726206	0.897955009
lecture_3	790	0.813915202	0.676678838	0.888194005	0.769372095	0.860734152	0.783003378	0.956836458	0.834428686	0.896615558
lecture_4	800	0.806765275	0.672456679	0.879546135	0.761860189	0.848619822	0.772944778	0.95973857	0.826269698	0.892833154
lecture_5	810	0.806114968	0.668931507	0.875569664	0.757684252	0.849752193	0.766579803	0.946929357	0.820723644	0.882956485
lecture_6	830	0.804543564	0.672798923	0.880405054	0.760743296	0.855795316	0.773682536	0.958727657	0.819620152	0.886444588
lecture_hfrp	818	0.904157149	0.772415401	0.980641699	0.835469267	0.93660989	0.864104636	0.986292481	0.900734148	0.982749692
lecture_hhp	236	0.948386261	0.790875782	0.991012903	0.896860465	0.952936069	0.875332296	0.990093	0.935033288	0.99051331
meeting_hfrp	337	0.919411336	0.777246634	0.980598222	0.858566405	0.944139205	0.874003824	0.984286396	0.916746526	0.983950967
meeting_hhp	254	0.948417357	0.789931825	0.991228907	0.895150171	0.952845166	0.872635935	0.990638349	0.935449713	0.989624313
office_hfrp	390	0.861631339	0.739781868	0.94723186	0.80439343	0.908749904	0.845520786	0.966303776	0.870839979	0.95720007
office_hhp	520	0.923050245	0.770420087	0.987552094	0.857421488	0.944508922	0.874508771	0.990008005	0.922724567	0.988547912

Chroma Recall

	RT60	YIN	TWM	MEL	SAC	YIN NML	TWM NML	MEL NML	SAC NML	MIX
dry	0	0.954254766	0.972609127	0.907535954	0.974653782	0.9492914	0.978360115	0.851382009	0.975676256	0.90873983
booth_1	80	0.940794071	0.968227235	0.902142914	0.962694461	0.93253005	0.9702018	0.826570622	0.959646078	0.890811803
booth_2	110	0.936651726	0.960592952	0.911155644	0.970236122	0.930693155	0.965953484	0.83461376	0.96547342	0.906421345
booth_3	180	0.929790721	0.946256854	0.918156055	0.958179746	0.933047168	0.9640282	0.856404934	0.965301384	0.921020665
meeting_1	210	0.934652041	0.962908262	0.914197438	0.968722156	0.930475666	0.967364919	0.810316874	0.966743621	0.911638318
meeting_2	220	0.932441702	0.95904229	0.919484621	0.966352107	0.923634849	0.966107292	0.817162161	0.967348433	0.915370456
meeting_3	240	0.930421496	0.957835887	0.91630345	0.964429419	0.923619217	0.967424662	0.815374363	0.957610095	0.909083891
meeting_4	250	0.92395998	0.9533059	0.914329895	0.960250919	0.92606574	0.960128963	0.814230112	0.962769707	0.909967977
office_1	370	0.908155845	0.942800484	0.908994519	0.954230611	0.908222984	0.962542069	0.794410896	0.964191316	0.910398364
office_2	440	0.886009776	0.916822857	0.895175922	0.929487962	0.882164494	0.941575981	0.798661866	0.937807493	0.894905736
office_3	480	0.866247952	0.892787564	0.881455723	0.909996837	0.864416315	0.926439746	0.750068938	0.929279475	0.881604787
lecture_1	700	0.90630488	0.931518796	0.919337666	0.950313252	0.918840417	0.959657641	0.823855482	0.963074716	0.924248591
lecture_2	720	0.875514648	0.90133022	0.894279944	0.920383669	0.887163436	0.936704093	0.783024806	0.942112435	0.903732287
lecture_3	790	0.868760155	0.89797454	0.889800352	0.923340654	0.869602748	0.932109753	0.761646806	0.928957164	0.880018298
lecture_4	800	0.861709406	0.891574663	0.887357506	0.918504225	0.869340533	0.920451237	0.752786338	0.931573529	0.890875388
lecture_5	810	0.859955605	0.886691097	0.882053727	0.912432418	0.860799478	0.915052852	0.741543672	0.921899521	0.884078098
lecture_6	830	0.864516041	0.88626647	0.883364487	0.914515134	0.86662785	0.918113566	0.760464837	0.924738845	0.87865454
lecture_hfrp	818	0.942472788	0.971008473	0.907361685	0.973174725	0.941531744	0.974661998	0.81053772	0.969490535	0.897253159
lecture_hhp	236	0.955529109	0.979462745	0.910797189	0.9779213	0.95360804	0.979533186	0.862533514	0.974541554	0.906613398
meeting_hfrp	337	0.944722522	0.968783641	0.911228141	0.973059175	0.940703877	0.973446012	0.819423627	0.973114013	0.902262033
meeting_hhp	254	0.955557188	0.97978055	0.910688144	0.9781667	0.95426524	0.979083346	0.862118898	0.974828731	0.908214162
office_hfrp	390	0.916491904	0.951193352	0.900529599	0.956627892	0.910934232	0.960958997	0.75808143	0.961109875	0.885520201
office_hhp	520	0.952675357	0.976439269	0.91871148	0.976673301	0.953045678	0.980922194	0.865495228	0.976843945	0.915928292

Bibliography

[1] E. Gómez, A. Klapuri, and B. Meudic, "Melody description and extraction in the context of music content processing," *Journal of New Music Research*, 2003.

[2] W. Hess, *Pitch determination of speech signals. Algorithms and devices.* Berlin, New York, Tokyo: Springer-Verlag, 1983.

[3] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Transactions on Signal Processing*, vol. 39, no. 1, pp. 40–48, 1991.

[4] D. Talkin, "Robust algorithm for pitch tracking," in in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. Elsevier Science B. V., 1995.

[5] A. Klapuri, "Qualitative and quantitative aspects in the design of periodicity estimation algorithms," *Proceedings of the European Signal Processing Conference*, 2000.

[6] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, p. 1917, 2002.

[7] E. Terhardt, "Calculating virtual pitch," *Hearing research*, vol. 1, pp. 155–182, 1979.

[8] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *The Journal of the Acoustical Society of America*, vol. 71, pp. 679–688, 1982.

[9] B. Gold and L. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *The Journal of the Acoustical Society of America*, vol. 46, pp. 442–448, 1969.

[10] L. Rabiner and R. Schafer, *Digital processing of speech signals.* Prentice-Hall, 1979.

[11] A. Bregman, "Psychological data and computational auditory scene analysis," in in *Computational auditory scene analysis*, D. Rosenthal and O. HG, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., 1998.

[12] A. Noll, “Cepstrum pitch determination,” *The journal of the acoustical society of America*, vol. 41, pp. 293–309, 1967.

[13] M. Lahat, R. Niederjohn, and D. Krubsack, “A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 6, pp. 741–750, 1987.

[14] M. Piszczalski and B. Galler, “Predicting musical pitch from component frequency ratios,” *The Journal of the Acoustical Society of America*, vol. 66, pp. 710–720, 1979.

[15] X. Rodet and B. Doval, “Fundamental frequency estimation using a new harmonic matching method,” *Proceedings of the International Computer Music Conference*, pp. 555–558, 1991.

[16] R. Maher and J. Beauchamp, “Fundamental frequency estimation of musical signals using a two-way mismatch procedure,” *The Journal of the Acoustical Society of …*, vol. 95, no. April, pp. 2254–2263, 1994.

[17] T. Jehan, “Musical signal parameter estimation,” *CNMAT report*, 1997.

[18] A. Klapuri, T. Virtanen, and J. Holm, “Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals,” *Proc. of the International Conference on Digital Audio Effects*, 2000.

[19] E. Pollastri, “A pitch tracking system dedicated to process singing voice for music retrieval,” *Multimedia and Expo, 2002. ICME’02. Proceedings. …*, pp. 9–12, 2002.

[20] C. Hsu and D. Wang, “A Tandem Algorithm for Singing Pitch Extraction and Voice Separation From Music Accompaniment,” *Audio, Speech, and …*, vol. 20, no. 5, pp. 1482–1491, 2012.

[21] J. Salamon and E. Gómez, “Melody extraction from polyphonic music signals using pitch contour characteristics,” *Audio, Speech, and Language …*, vol. 20, no. 6, pp. 1759–1770, 2012.

[22] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, p. 3690, 2004.

[23] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," *Proceedings of ICASSP*, pp. 756–759, 1995.

[24] P. Naylor and N. Gaubitch, *Speech dereverberation*. Springer, 2010.

[25] Y. Haneda, S. Makino, and Y. Kaneda, "Common acoustical pole and zero modeling of room transfer functions," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 320–328, 1994.

[26] J. Hopgood and P. Rayner, "A probabilistic framework for subband autoregressive models applied to room acoustics," *Proc. IEEE Workshop Statistical Signal Processing*, pp. 492–495, 2001.

[27] J. Hopgood and P. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 476–488, 2003.

[28] J. Mourjopoulos and M. Paraskevas, "Pole and zero modeling of room transfer functions," *Journal of Sound and Vibration*, vol. 146, no. 2, pp. 281–302, 1991.

[29] J. Mourjopoulos, "Digital equalization of room acoustics," *Journal of the Audio Engineering Society*, vol. 42, no. 11, pp. 884–900, 1994.

[30] T. Paatero, "Modeling of long and complex responses using Kautz filters and time-domain partitions," *Proc. European Signal Processing Conf.(EUSIPCO)*, pp. 313–316, 2004.

[31] H. Kuttruff, *Room acoustics*, 4th ed. Taylor & Francis, 2000.

[32] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of acoustics*, 4th ed. John Wiley and Sons, Inc., 2000.

[33] W. T. Chu, "Comparison of reverberation measurements using schroeders impulse," *Journal of Acoustical Society of America*, vol. 63, pp. 1444–1450, 1978.

[34] J. Jot, L. Cerveau, and O. Warusfel, “Analysis and synthesis of room reverberation based on a statistical time-frequency model,” *Proc. Audio Eng. Soc. Convention*, 1997.

[35] M. Miyoshi and Y. Kaneda, “Inverse filtering of room acoustics,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, 1988.

[36] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, “Speech dereverberation based on variance-normalized delayed linear predictor,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[37] A. Krueger and R. Haeb-Umbach, “A model-based approach to joint compensation of noise and reverberation for speech recognition,” in in *Robust Speech Recognition of Uncertain or Missing Data: Theory and Applications*, D. Kolossa and R. Haeb-Umbach, Eds. Springer-Verlag, 2011, pp. 257–290.

[38] S. Haykin, *Adaptative Filter Theory*, 4th ed. Prentice-Hall, 2011.

[39] B. Gillespie and A. Atlas, “Strategies for improving audible quality and speech recognition accuracy of reverberant speech,” *Proc. Int. Conf. Acoust., Speech, Signal Process.*, pp. 676–679, 2003.

[40] M. Triki and D. Slock, “Delay and predict equalization for blind speech dereverberation,” *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 11, no. 5, pp. 97–100, 2006.

[41] H. Buchner and W. Kellermann, “TRINICON for dereverberation of speech and audio signals,” in in *Speech Dereverberation*, P. Naylor and N. Gaubitch, Eds. Berlin: Springer-Verlag, 2010, pp. 311–385.

[42] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, “Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction,” *Audio, Speech, and …*, vol. 17, no. 4, pp. 534–545, 2009.

[43] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. Okuno, “Blind separation and dereverberation of speech mixtures by joint optimization,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 69–84, 2011.

[44] E. Habets, “Single- and multi-microphone speech dereverberation using spectral enhancement,” Eindhoven Univ. Technology, 2006.

[45] J. Erkelens and R. Heusdens, “Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1746–1765, 2010.

[46] T. Y. Hirokazu Kameoka, Tomohiro Nakatani, “Robust speech dereverberation based on non-negativity and sparse natures of speech spectrograms,” *Acoustics, Speech and …*, 2009.

[47] K. Kumar, R. Singh, B. Raj, and R. Stern, “Gammatone sub-band magnitude-domain dereverberation for ASR,” *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, no. 1, pp. 4604–4607, May 2011.

[48] K. Kumar, B. Raj, R. Singh, and R. M. Stern, “An iterative least-squares technique for dereverberation,” *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, no. 2, pp. 5488–5491, May 2011.

[49] K. Lebart, J. Boucher, and P. Denbigh, “A new method based on spectral subtraction for speech dereverberation,” *Acta Acustica united with Acustica*, vol. 87, pp. 359–366, 2001.

[50] D. Wang and G. Brown, “Computational auditory scene analysis: Principles, algorithms, and applications,” in in *Chapter 7.3*, 2006.

[51] “MIR-1K webpage.” [Online]. Available:
https://sites.google.com/site/unvoicedsoundseparation/mir-1k. [Accessed: 28-Aug-2013].

[52] C. Hsu and D. Wang, “A tandem algorithm for singing pitch extraction and voice separation from music accompaniment,” *Audio, Speech, and …*, 2012.

[53] “MIR-1K for MIREX: pitch interpolation.” .

[54] “MIREX 2012 audio melody extraction.” .

[55] "MIREX 2005 audio melody extraction results." [Online]. Available: http://www.music-ir.org/mirex/wiki/2005:Audio_Melody_Extraction_Results. [Accessed: 28-Aug-2013].

[56] D. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," *Journal of Machine Learning Technologies*, 2011.

[57] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," *Digital Signal Processing, 2009 …*, 2009.

[58] M. Jeub, M. Schäfer, and H. Krüger, "Do we need dereverberation for hand-held telephony?," *Proc. 20th Int. …*, no. August, pp. 1–7, 2010.

[59] "Aachen Impulse Response dataset website." [Online]. Available: http://www.ind.rwth-aachen.de/en/research/tools-downloads/aachen-impulse-response-database/. [Accessed: 29-Aug-2013].

[60] M. Schroeder, "New Method of Measuring Reveberation time," *The Journal of the Acoustical Society of America*, 1965.

[61] R. Lyon, "History and future of auditory filter models," *Circuits and Systems (…*, 2010.

[62] K. Kinoshita, T. Nakatani, M. Miyoshi, and T. Kubota, "A new audio postproduction tool for speech dereverberation," *Watermark*, 2008.

[63] D. Gesbert and P. Duhamel, "Robust blind channel identification and equalization based on multi-step predictors," *Acoustics, Speech, and Signal …*, pp. 0–3, 1997.

[64] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *Acoustics, Speech and Signal Processing, IEEE …*, no. 2, pp. 113–120, 1979.

[65] K. Kinoshita, "Spectral subtraction steered by multi-step linear prediction for single channel speech dereverberation," *Acoustics, Speech and …*, pp. 817–820, 2006.

[66] K. Kinoshita and M. Delcroix, "A linear prediction-based microphone array for speech dereverberation in a realistic sound field," *Proc. of Audio …*, 2007.