PITCH-TRACKING OF REVERBERANT SOUNDS, APPLICATION TO SPATIAL DESCRIPTION OF SOUND SCENES

ALEXIS BASKIND AND ALAIN DE CHEVEIGNÉ

IRCAM, 1 place Igor-Stravinsky, 75004 Paris, France
Alexis.Baskind@ircam.fr
Alain.de.Cheveigne@ircam.fr

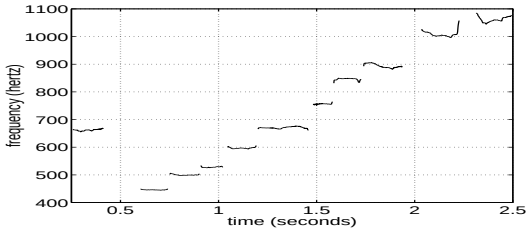
Fundamental frequency (F0) is useful as a perceptually-relevant sound descriptor, and also as an ingredient for signal processing applied to analysis of sound scenes. Here, a recently proposed multiple-F0 algorithm is adapted to handle reverberation in monophonic or multichannel recordings; the information that is obtained from it is then applied to estimation of reverberation time from recorded musical signals.

INTRODUCTION

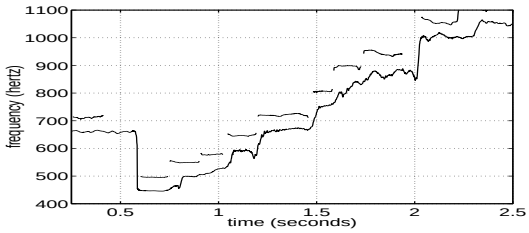
Fundamental frequency ("F0") estimation is an initial step in many systems for the analysis of complex sound scenes, such as speech recognition, score following, low-bitrate coding of musical signals, etc... Many algorithms had been developped for this purpose, the great majority of them relying on time-frequency or time-lag analysis. Most of those techniques require the assumption of a single periodic signal at each instant, and thus are designed for monodic signals. Their behavior in the presence of background noise depends mainly on the signal-to-noise ratio and the decorrelation between signal and noise, the latter being often assumed stationnary. A recent pitch-tracker, called YIN, has proven to be robust and efficient, and is also fast enough to be implemented in real-time [8].

The presence of reverberation makes the F0 estimation task more difficult, as its spectral structure competes with that of the direct sound. Thus most of pitch-tracking devices fail at estimating the fundamental frequency of reverberant sounds with good accuracy, especially at transients. Figure 1 shows an example of this breakdown, for the very first seconds of a recording of Jean Sebastien Bach's "partita" for solo flute: at top is drawn the estimation provided by YIN on the dry recording, and at bottom the estimation provided by YIN on a reverberant version of this recording, synthesized by convolution with an artificial impulse response which reverberation time at low frequencies is approximately 1.5 second, and which clarity index C_{80} is +6 dB. What can be easily seen is a blurring of the estimation, especially when notes are close to each other in time and/or frequency. This sluggishness, which is undoubtedly due to the presence of both current direct sound and reverberation of the preceding notes, is a great disturbance for any kind of further analysis that requires an accurate estimation of running fundamental frequency.

A closer look at reverberation gives a clue to overcome this problem: since the reverberant tail is made of the superimposition of partially coherent echoes of the direct sound, the periodicity of the latter provides a strong constraint on the spectral content of the former. Although reverberation associated to a harmonic signal cannot be considered globally as strictly harmonic[1], the assumption of local harmonicity remains reasonnable (see figure 2). The aim of this study is to provide a fundamental-frequency estimator suitable for reverberant monodic recordings, that



(a) single-f0 estimation of the dry signal



(b) single-f0 estimation of the reverberant signal (dry reference is plotted in light gray)

Figure 1: single-f0 estimation of monodic music, in dry (top) and reverberant (bottom) conditions

[1] Actually, inharmonicity of the reverberant decay seems to be a relevant cue for estimating reverberation time [15].

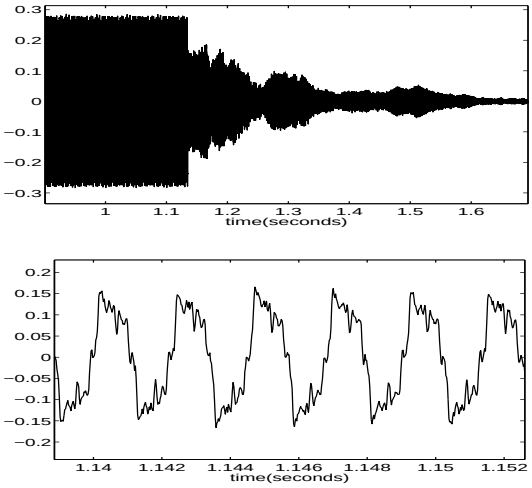


Figure 2: Global shape (top) and detail (bottom) of a reveberated square wave (f0=440 Hz)

takes into account the specific behavior of such signals. For that purpose, we adapt a recently-developped multiple-F0 estimation method called MMM [9], which is based on YIN, to the task of estimating the F0 of the direct and reverberated parts of a monodic recording. MMM works by jointly cancelling the various harmonic sources present, by searching through a two-dimensional lag space. The coordinates of the minimum give the two estimated periods of the signal, one being assigned to the direct sound, the other to the reverberation. Estimation is made more reliable by constraining the reverberated F0 to be within the range of recent values of the direct F0, and also by working on several channels.
F0 estimates, once obtained, are used to tune comb filters to isolate successive streams one from another, in order to perform analysis of the spatial features of the scene. As a detailed example of application, a method for estimating reverberation time from musical signals is proposed here, which is based on the derivation of short-time pitch-synchronous spectra of such isolated streams. An analysis of the decay is performed on each channel of those spectra between time limits that are defined with the knowledge of interchannel pitch-synchronous short-time coherence.
Knowledge of reverberation characteristics, as well as other spatial features of the scene, is of use for many applications related to production and post-production of multichannel sound (such as automated mixing, or cinema dubbing), or to indexation of binaural and multichannel recordings in databases [5].

1. DOUBLE-F0 ESTIMATION FOR REVERBERANT MUSIC

1.1. Cancellation model for double-f0 estimation

Like YIN, MMM relies on a cancellation model of pitch perception [7], which is an alternative to Licklider's traditionnal autocorrelation model [12]. Both are physiologically plausible, and have many similarities. However, as it will be shown, cancellation model has some quite interesting features for our purpose, since it is an intuitive point of view providing data that can directly be interpreted as a cue for judging the quality of the estimation.
The principle of double cancellation may be illustrated by diagram in figure 3.

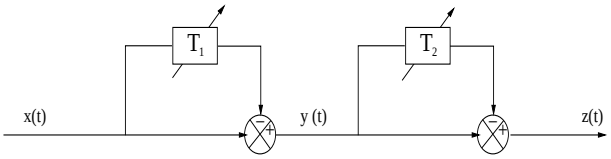


Figure 3: Double cancellation model

The principle is the following: Considering the instantaneous power $P_x(t)$ of a signal $x(t)$:

$$P_x(t) = \sum_{i=t}^{t+W} x^2(i)$$

(W is the length of the window), as well as the signal $z(t)$ that results from double comb-filtering by lags T_1 and T_2:

$$z(t) = x\,(t) - x\,(t - T_1) - x\,(t - T_2) + x\,(t - T_1 - T_2)$$

, the algorithm looks for the lags T_1 and T_2 that cancel this residue $z(t)$ the best, by minimizing its power, called **double difference function** (d.d.f.):

$$ddf\ (t, T_1, T_2) = P_z(t)$$

The d.d.f. is thus a running bi-dimensional pattern, which depends on the two internal variables, that is the lags used in this cancellation model. An example of the time evolution of this pattern is provided on figure... The couple of lags that minimize the d.d.f. are thus the estimations of the periods of the two harmonic sources that are assumed to be mixed. The quality of the estimation, as well as the relevance of this assumption, can be evaluated thanks to the following ratios, called "aperiodicity measures", all bounded between 0 and 1:

$$ap(t) = P_z(t)/P_x(t)$$

,

$$ap_1(t) = P_z(t)/P_{y_1}(t)$$

and

$$ap_2(t) = P_z(t)/P_{y_2}(t)$$

,

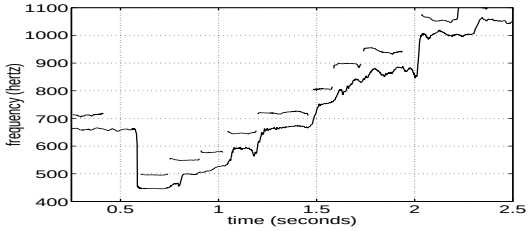
where $y_1(t)$ and $y_2(t)$ are the signals that result from a single cancellation, i.e. $y_1(t) = x(t) - x(t - T_1)$ and $y_2(t) = x(t) - x(t - T_2)$. The first ratio $ap(t)$ allows to evaluate the quality of the double-f0 model taken as a whole, whereas $ap_1(t)$ and $ap_2(t)$ are useful to compare the estimation to the single-f0 estimation that YIN provided. All this data, added to YIN data, is processed by a decision module, which decides which of the models (i.e. one or two harmonic sources) is the most accurate, and what is (or are) the fundamental frequency(ies) of the corresponding source(s).

1.2. The case of reverberant sounds

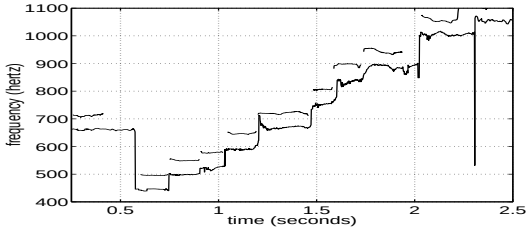
Applied directly to our specific concern, which is reverberant and monodic music, this algorithm is not fully satisfactory: whereas it works efficiently for two actual harmonic sources mixed together, it fails at detecting the reverberation of a single source. The main reason is the relative lack of harmonicity of the reverberant tails: since the signal contains an actual harmonic source, the algorithm often tends to divide it in two components, so that the decision module tends to choose the single-F0 model as more relevant than the double-F0 model. In those cases, the reverberant stream is thus not detected.
However, this problem can be overcome, taking into account the fact that the running fundamental frequency of the reverberant tail of a sound (assuming local harmonicity) is quite close to the actual fundamental frequency of the sound itself. Thus, by constraining one of the estimations to rely within bounds that would be determined by the main F0 estimation in the very near past (several hundreds of milliseconds), we expect the algorithm to detect the reverberant stream more efficiently. (see figure 4).

Figure 4: Double-F0 estimation for reverberant monodic sounds. Frequency bounds are determined by the prominent frequency in the last 300 ms.

As a practical example, is shown the result of the double-F0 estimation of the same short excerpt as in introduction, and comparison with single-f0 estimation: transients are precisely detected, thus most of notes can be distinguished anew as discrete events with a nearly constant fundamental frequency.
In the case of two-channel or multichannel recordings, we can also benefit from the redundant information on fundamental frequency over all recording channels that contain enough direct sound. Different approaches could be employed in order to use this additional information. Ours at present relies on the same overall principle as the single channel estimation, but whereas a separate single- and double-F0 estimation is performed on each channel, the decision module is common to all channels, providing a unique fundamental frequency estimation at a given time, corresponding to the lowest residual aperiodicity. It is also worth noticing that the cancellation model, when generalized to two or several channels, is suitable for estimating the delays **between channels**. Thus, a possible extension of this architecture would be for instance a binaural signal detector, which could at the same time estimate the localization of the source (or at least its lateralization) and its pitch.



(a) single-f0 estimation of the dry signal (dry reference is plotted in light gray)



(b) single-f0 estimation of the reverberant signal (dry reference is plotted in light gray)

Figure 5: comparison between single-f0 (top) and double-f0 (bottom) estimation of monodic reverberant music

2. APPLICATION TO RUNNING ESTIMATION OF REVERBERATION TIME

2.1. The problem of estimating reverberation time from music

The idea of deriving reverberation time from musical signals is not new, since it is one of the ways to solve the major problem of calculating RT in occupied halls. As a matter-of-fact, traditional impulse response measurements, using pseudo-random noises or short impulses such as gun shots, cannot be used in this case, which correspond most of the time to the situation of a concert. As an alternative to predictive estimations from measurements made in the empty hall [3, 6], many acousticians tried to derive reverberation time directly from the music that is diffused during the concert, focusing for instance on differences between modulation transfer functions measured close to the musicians and in the audience [13], ob-

serving the shape of the autocorrelation envelope [10], or analyzing decays after "stop-chords" in the music [6]. This latter method, which does not need the knowledge of the source signal, may be very useful (and also very sensitive to the "quality" of the silence [5]), but what can be done if there is not any complete silence during the whole concert ?

The solution that is proposed here is inspired from our audition: of course, we are able to hear late reverberation during complete silences, but also just after sudden frequency changes, when the source is narrowband or harmonic. Actually, both situations provide decays which may be uncorrupted by a direct sound during a sufficient time period.

A useful application of the estimations provided by the pitch-tracker presented above can thus be foreseen at that point: assuming that the fundamental frequency of the source signal as a function of time is known with good accuracy, we get a quite strong information on the sound scene, that could be used at least in two different (and complementary) ways: first, the frequency bands in which reverberation actually occurs are known, since they correspond to the fundamental frequency of the present note and its harmonics; second, the fundamental frequencies of the past and future notes can be directly used to cancel them the best possible, in order to clean the decay from outer disturbances. This latter operation, which could be performed thanks to basic first-order comb filters, is a major help in our attempt to isolate the decay.

2.2. Pitch-synchronous time-frequency analysis of reverberant decays

The method for estimating reverberation time that is envisionned here must rely on an accurate time-frequency front-end analysis. Using short-time Fourier transform with a constant number of frequency channels is convenient and allows the use of the well-known FFT optimized algorithm, but since the instantaneous pitch of the direct sound is known, it is possible to achieve a better precision by using pitch-synchronous analysis. The principle remains the same, except that the number of frequency channels now depends on the fundamental frequency of the signal, so that any of them match a harmonic of the note.

Short-time Fourier spectrum

Short-time Fourier spectrum is often used as a basic ingredient for describing reverberation in narrow bands [11], since it is a fast, intuitive and convenient method for time-frequency analysis. It has been choosen here mainly because of it allows pitch-synchronous analysis, but may be replaced with success by other techniques, such as modified discrete cosine transform, as well as constant-Q or ERB filterbanks.

Applied directly on the signal, short-time Fourier spectrum rarely reveals decays during a period that is sufficient for further analysis, since even in the case where the following notes do not share the same frequency bands as the present note, their presence remains visible in the pitch-synchronous spectrum, mainly because the band-pass of the analysis window is too large. That is why the signal is first preprocessed in order to reduce those disturbances. This preprocessing just consists in several comb-filters which delays match the periods of the preceding and following notes. On figure 6 is provided an example of the performance of this quite simple method: the considered reverberated note of this flute recording, which mean F0 is 446 Hz (i.e. an 'A'), just follows the first 'E' (662 Hz) of the melody, and is directly followed by a 'B' (499 Hz) and a 'C' (527 Hz). All of them are to be cancelled by this preprocessing stage. The comparison of the spectra at the fundamental frequency shows that cancelling allowed to reveal the onset of the note (arround 0.2 seconds), as well as the exponential decay (between 0.2 and 0.5 seconds). However it is also easily visible that every additional comb-filtering tends to lower the signal to noise ratio; thus we have to be very careful not to apply too many filters.

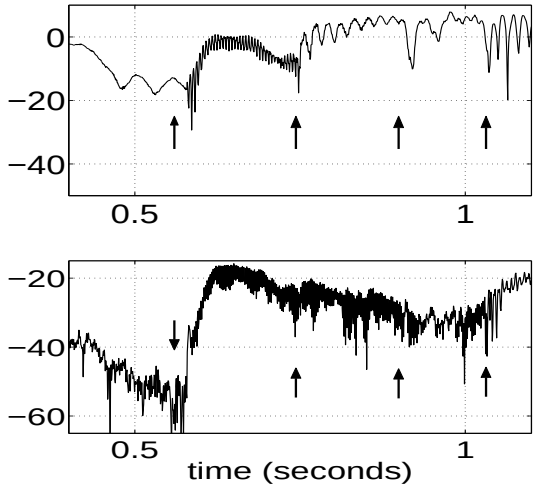


Figure 6: Example of the efficiency of pitch-synchronous comb-filtering: energy within the fundamental frequency band. Arrows indicate onset times of actual and following notes.
Top: without filtering preceding and following notes. *Bottom*: with filtering

However, the precision that may be achieved by directly applying a linear regression method over short-time Fourier spectrum segments is quite poor, especially at low frequencies. This is due to oscillations in the decay that correspond to the stochastic nature of reverberation. A well known method to reduce this variance is backward integration of instantaneous power. This technique, that has been proposed by Schroeder [14] in order to reduce

the number of measurements that are needed to derive reverberation time, was designed for octave or third-octave band-filtered impulse responses, but has been applied with success to narrow-band representations [11, 4]. Applying it on segments of musical signals instead of impulse responses makes sense, but entails some additional difficulties that are related to two reasons: first, the possible presence of background noise during the estimation, which is most of time not stationary since it corresponds to other sources and to reverberation; second, the necessity to define accurately times for the beginning and the end of the analysis. Both problems can be handled efficiently thanks to an additional cue, short-time coherence.

Short-time interchannel coherence

- Coherence a court-terme => marqueurs => régression => voir [1], et [2]

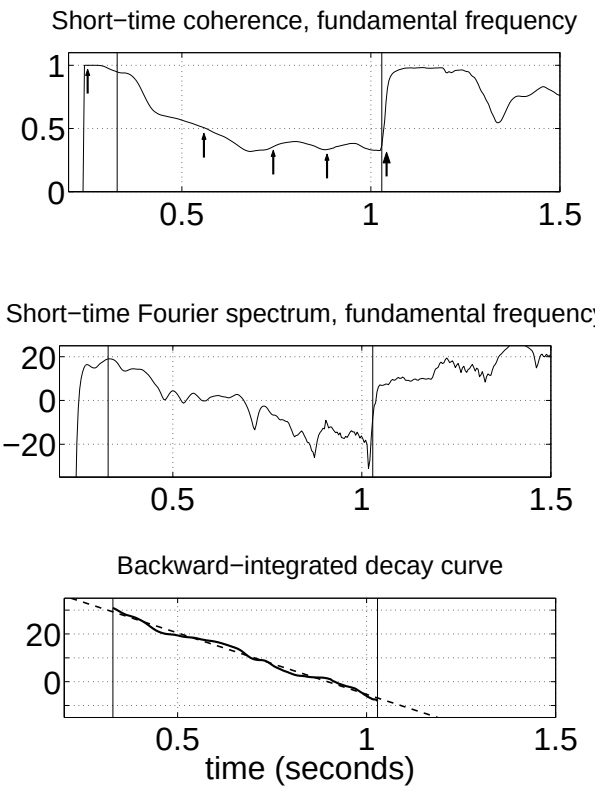


Figure 7: The principle of running estimation of reverberant decay on a specific example. All plots concern the frequency channel corresponding to fundamental frequency of the actual note (662 Hz). Vertical lines are calculated limits for regression. Arrows on top figure are onset times of actual and following notes.
Top: short-time coherence. *Middle*: short-time Fourier spectrum. *Bottom*: backward integration of Fourier spectrum

2.3. Practical example

This method is applied on a longer excerpt (22 seconds) of the example used above: the dry recording that has been used, which is a recording solo flute close to the musician, has been spatialized by convolution with a set of binaural room impulse responses synthesized by IRCAM's "Spatialisateur", with the following objective measurements: clarity index C_{80} equals 6.05 dB, and in the frequency range of interest, early decay time EDT_{10} equals approximately 1 second, and RT_{30} equals some 1.5 seconds. The two-channel "binaural" recording that is obtained is processed first by the pitch-tracker, and then by the reverberation time estimation device. Among the 330 notes of this recording, 111 where choosen for their sufficient signal-to-noise ratio, providing 120 narrow-band segments that match the requirements on coherence and dynamic range.

All the estimates were gathered in third-octave bands. For each band, the mean value as well as the standard deviation is computed. The standard deviation provides an estimation of the "confidence interval" (this term is not fully accurate, since the estimation for each band is not gaussian). On figure8 are shown the results of this analysis, as well as the actual reverberation time RT_{30} and early decay time EDT_{10}, computed with Brian Katz' "Impulse Response Analysis" toolbox (those measurements are coherent with narrow-band measurements provided by "EDR", IRCAM room acoustics team's toolbox for room analysis).

If we do not consider the higher bands, which estimations are not at all correct, mainly becauser of the lack of relevant data, it is obvious from this plot that most of the estimated decay times are comprised between early decay time and reverberation time. A closer look at the data shows that the longer a given segment is, the better the estimation matches the reverberation time, and the shorter this segment is, the better the estimation matches the early decay time. Those results are consistent with physics and perception of reverberation: as a matter-of-fact, running reverberance, to which EDT corresponds, is audible anytime during the signal as soon as it is not continuous, and late reverberance, which is better described by RT, is audible only during silences or sudden pitch changes.

CONCLUSION

This paper presents methods that aim at deriving two perceptually-relevant objective descriptors of a reverberant sound scene: the pitch, and the reverberation time. The pitch-tracker that is proposed here provides very encouraging results; the estimation itself, as well as the decision module are constantly improved in order to provide more accurate results, but it is at this stage anyway better suitable for reverberant signals than usual single-voice pitch-trackers. The method for deriving reverberation time that was de-

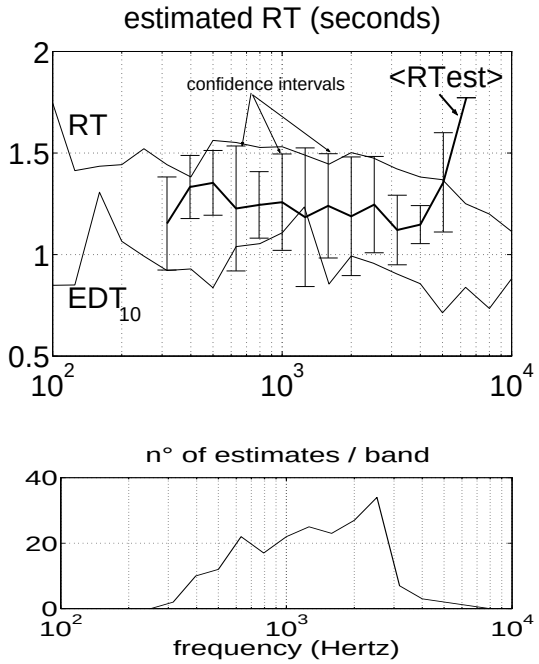


Figure 8: Results of the estimation over a 22 seconds excerpt. *Top*: mean estimation and "confidence intervals", compaired with actual RT_{30} and EDT_{10}. *Bottom*: number of estimates in each third-octave band

velopped shows remarkable adequacy to traditional objective measurements, even if it is not at present able to distinguish actual reverberation time from early decay time. Further work is achieved in that direction, which mainly consists in finding the adequate configuration for each purpose.

The use of the fundamental frequency information in deriving a spatial description is not limited to the estimation of reverberation time; it provides a strong information for a further complete analysis of all spatial features of a sound scene, such as for instance onset detection and localization.

REFERENCES

[1] J. Allen, D. Berkley, and J. Blauert. Multimicrophone signal processing technique to remove room reverberation of speech signals. *J. Acoust. Soc. Am.*, 62(2):912–915, 1977. 2.2

[2] Carlos Avendano and Jean-Marc Jot. Frequency domain techniques for stereo to multichannel upmix. In *Proc. AES 22nd international conference, "Virtual, synthetic and entertainment audio", Espoo, Finland*, pages 121–130, June 2002. 2.2

[3] M. Barron. *Auditorium acoustics and Architectural Design*. E & FN Spon/Chapman & Hall, 1993. 2.1

[4] A. Baskind and J.-D. Polack. Sound Power Radiated by Sources in Diffuse Field. In *proc. AES 108th convention*, February 2000. 2.2

[5] Alexis Baskind and Olivier Warusfel. Methods for blind computational estimation of perceptual attributes of room acoustics. In *proc. AES 22nd international conference on virtual, synthetic and entertainment audio*, Espoo, Finland, June 2002. 1, 2.1

[6] Leo L. Beranek. *Concert and opera halls : how they sound*. Acoustical Society of America, 1996. 2.1

[7] Alain de Cheveigné. Cancellation model of pitch perception. *J. Acoust. Soc. Am.*, 103(3):1261–1271, march 1998. 1.1

[8] Alain de Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917–1930, April 2002. (document)

[9] Alain de Cheveigné and Alexis Baskind. F0 estimation of one or several voices. In *Proc Eurospeech (submitted)*, 1993. 1

[10] Martin Hansen. A method for calculating reverberation time from musical signals. Technical report, report from the Acoustics Laboratory, Technical University of Denmark, 1995. Report no 60, ISSN 0105-3027. 2.1

[11] J.-M. Jot, L. Cerveau, and O. Warusfel. Analysis and synthesis of room reverberation based on a time-frequency model. In *AES 103rd convention preprint*. AES, September 1997. 2.2, 2.2

[12] J.C.R Licklider. A duplex theory of pitch perception. *Experientia*, 7(4):128–132, 1951. 1.1

[13] J.-D. Polack, H. Alrutz, and M. R. Schroeder. The modulation transfer function of music signal and its applications to reverberation measurement. *Acustica*, 54:257–265, 1984. 2.1

[14] M. R. Schroeder. New method for measuring reverberation time. *J. Acoust. Soc. Am.*, 37:409–412, 1965. 2.2

[15] M. Wu and D. Wang. A one-microphone algorithm for reverberant speech enhancement. To be presented in ICASSP2003, Hong Kong, April 6-10, 2003, 2003. 1