Harmonic Sinusoidal+Noise Modeling of Audio based on Multiple F0 Estimation

Maciej Bartkowiak, Tomasz Żernicki

Poznań University of Technology, Chair of Multimedia Telecommunications and Microelectronics,
Polanka 3, 60-965 Poznań, Poland
mbartkow@multimedia.edu.pl, tzernicki@multimedia.edu.pl

ABSTRACT

This paper deals with the detection and tracking of multiple harmonic series. We consider a bootstrap approach based on prior estimation of F0 candidates and subsequent iterative adjustment of a harmonic sieve with simultaneous refinement of the F0 and inharmonicity factor. Experiments show that this simple approach is an interesting alternative to popular strategies, where partials are detected without harmonic constraints, and harmonic series are resolved from mixed sets afterwards. The most important advantage is that common problems of tonal/noise energy confusion in case of unconstrained peak detection are avoided. Moreover, we employ a popular LP-based tracking method which is generalized to dealing with harmonically related groups of partials by using a vector inner product as the prediction error measure. Two alternative extensions of the harmonic model are also proposed in the paper that result in greater naturalness of the reconstructed audio: an individual frequency deviation component and a complex narrowband individual amplitude envelope.

1.   INTRODUCTION

Harmonic sinusoidal modeling first proposed by Serra [1] is a well established signal analysis and processing framework applicable to speech and musical sounds, mostly generated by individual instruments. Serra's hybrid approach assumed that a fundamental frequency f_0 of the given sound exists and is known, so that the signal may be well approximated as a non-stationary harmonic series and the residual noise,

$$\hat{x}(t) = \sum_{k=1}^{K} A_k(t) \sin\left(\varphi_k + 2\pi \int_0^t k\, f_0(\tau)\, d\tau \right) + r(t)\,. \quad (1)$$

Perhaps the most important motivation of introducing harmonic constraints to the sinusoidal model was to provide a criteria that allowed unambiguous discrimination between deterministic (tonal) and stochastic (noise-like) components of audio spectra. In general, the separation of these parts is a difficult problem. First of all, the bulk of spectral components observed in natural audio exhibit only certain degree of coherence in time evolution of phase and instantaneous

frequency. Consequently, most of them is neither purely deterministic nor purely random. In fact, the distinction alone is not as much critical from the perceptual point of view, as it is important due to the representation efficiency (in applications related to compression) and flexibility (in applications involving sound transformations).

Serra's hybrid harmonic model is limited to single, monophonic sounds. Harmonic modeling of polyphonic music is currently an important research problem and a hot industrial topic. Having a reliable harmonic modeling tool at hand would allow for ultimate sound manipulations, remixing of music records, source separation and recognition, not to mention a very efficient object-oriented representation and coding [2]. Our goal is to obtain an perceptually accurate representation of the polyphonic music signal so that a short frame of it may be expressed as

$$\hat{x}(t) = \sum_{k=1}^{K} \sum_{n=1}^{N_k} A_{n,k}(t) \sin\left(\varphi_k + 2\pi \int_0^t f_{n,k}(\tau)\, d\tau\right) + r(t) \quad (2)$$

where N_k denotes a number of components in each of the K harmonic series, and $f_{n,k}$ denote the frequencies harmonically related to the corresponding fundamentals, $f_{0,k}$.

Recent attempts at harmonic modeling of polyphonic audio may be categorized in two groups. The first group covers employing of harmonic matching pursuit aimed at detection of whole series of partials at once [3,4]. The important limitation of such methods besides their huge computational complexity is the inability to detect harmonic structures with varying fundamental frequencies. The approaches from the second group involve an application of harmonic constraints at the early stage of sinusoidal analysis, after spectral peak detection [5]. They require solving a set of equations governing given set of partial frequencies that are estimated previously without harmonic constraints. Within these methods, estimation of f_0 is based on an a'posteriori statistical reasoning. The main disadvantage of such approach is that the front-end unconstrained peak detection is prone to tonal/noise energy confusion and thus the seek of harmonic series may be performed on spurious data.

The approach proposed in this paper is quite different from the above two. We first attempt to pre-estimate candidate fundamental frequencies from the power

spectrum of the original signal, and on this basis we try to resolve sets of partials that are (almost) harmonically related to one of the multiple f_0's (fig. 1). The second stage of this algorithm is in fact an iterative optimization procedure aimed at finding spectral peaks that are not necessarily at the ideal overtone positions, but in the close proximity. This is achieved by adjusting the fundamental f_0 and β, so that the observed overtone frequencies f_n satisfy the model

$$\hat{f}_n \cong n\, f_0 \sqrt{1 + \beta\,(n^2 - 1)}\,. \quad (3)$$

The groups of harmonically related partials are tracked on a frame by frame basis. We employ a well known linear-prediction based tracking technique [6] which is modified to deal with vectors of partial frequencies instead of individual frequencies.
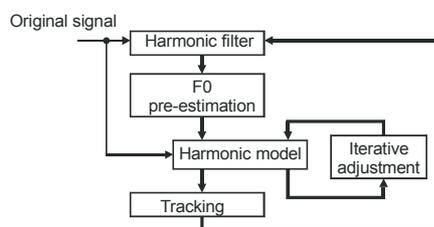

Figure 1. The general scheme of the proposed harmonic model based on pre-estimated fundamental frequencies

Furthermore, we propose two extensions of the classic harmonic model that in our experience lead to greater flexibility and higher quality of the reconstructed signals. The first extension is that each amplitude of an individual partial is allowed certain bandwidth and is represented by a narrowband complex signal instead of the piecewise linear segments commonly used in the classical approach. Such representation is more truthful, as it embraces the whole of the signal energy associated with given harmonic partial, including that related to frequency estimation error. It also takes into account the natural width of each spectral line coming from the already mentioned phase incoherence. Therefore our model is capable of representing small random fluctuations of amplitudes and frequencies observed on short-time spectrograms of natural sounds.

The second alternative extension is that each individual harmonic partial is allowed some minor frequency deviation from the (in)harmonic series (2-3). This yields a more natural and "live" sounding re-synthesized audio, without the artificial timbral characteristics

which is a typical effect of rigid overtone frequency relations.

The paper is organized as follows. A simple technique for estimation of multiple fundamental frequencies is presented in section 2. An iterative harmonic series estimation algorithm is proposed in section 3. Section 4 describes a harmonic tracking technique using a vector extension of the linear prediction based tracking. The extensions of the harmonic model are discussed in section 5. Some experimental results and a discussion is offered in section 6.

2. ESTIMATION OF MULTIPLE FUNDAMENTAL FREQUENCIES

A great number of f_0 estimation techniques has been proposed for analysis of speech and audio [7] using either a time-domain or frequency-domain approaches with mixed results on real-life data. Some of them, motivated by the success of the human auditory system, employ perceptual principles and scales for the detection of pitch that is closely related to f_0 [8]. Unfortunately, many of these techniques offer only a moderate performance in the case of polyphonic music containing harmonies, reverberation and noise components [7]. Multiple fundamental frequency estimation in music recordings is still considered as a difficult problem. On the other hand, it has been shown recently that the most reliable results of f_0 estimation are offered by taking into account a long term coincidental evolution of many harmonic partials (e.g. [9]). A sinusoidal model is a proper tool for revealing temporal organization in a complex audio signal. On the other hand, only harmonic model delivers partial parameters properly structured in harmonic groups. Such grouping, is much easier a task for the algorithm that is aware of fundamental frequencies. Therefore, a harmonic sinusoidal model may be employed together with f_0 estimation in a self-sustained estimation loop. Within such approach, f_0 estimation relies on the results of the harmonic model that in turn relies on the f_0 estimation (fig. 1). For a bootstrap initialization, a simple pre-estimation of f_0 candidates is necessary.

Our technique for multiple f_0 pre-estimation relies on two heuristic assumptions. First of all, it is assumed that a most prominent spectral peak in the signal short-time power spectrum is a member of the most prominent harmonic series (fig. 2, upper plot). In other words, the most prominent peak is either a fundamental or an overtone of an important fundamental that should be

detected in the first order. The second assumption is that other harmonic series exhibit significant spectral peaks besides those from the currently considered one, so that they still may be detected after removal of the partials of the already detected series from the signal spectrum. As it shall be demonstrated, even if these assumptions do not hold in all cases, a pre-estimation of f_0 on this basis is enough for the proper bootstrap harmonic modeling process.

The signal is analyzed in overlapping frames windowed by a Gaussian window and zero-padded. A power spectrum $S(f) = |X(f)|^2$ is calculated in each frame using FFT of a sufficiently high resolution. The most prominent peak is detected. The exact frequency of this peak, f_p is carefully determined by the use of a spectral estimation technique [10]. All integer sub-multiples f_c of this frequency down to the minimum allowed frequency f_{min} are calculated and considered as potential fundamental frequencies. For each considered f_c, a MOP (mean overtone power) descriptor is calculated,

$$MOP(f_c) = \frac{1}{N_c} \sum_{n=1}^{N_c} \int_0^{f_{max}} S(f)\, G_{nf_c}(f)\, df \qquad (4)$$

where $\quad G_{nf_c}(f) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\!\left(-\dfrac{(f - n f_c)^2}{2\,\sigma^2} \right),$

$$f_c = \frac{1}{c} f_p, \quad c = 1,2\ldots\left\lceil \frac{f_p}{f_{min}} \right\rceil,$$

$$N_c = \left\lfloor \frac{f_{max}}{f_c} \right\rfloor, \quad f_p = \arg \max_{f_{min} < f < f_{max}} |X(f)|^2$$

The value of the MOP descriptor is the average power of narrow spectral bands comprising all overtones of the candidate fundamental f_c. These are calculated by integrating the power of the signal spectrum weighted by appropriate normal pdf-s (4).

It is interesting to observe, how the value of the MOP descriptor depends on the candidate fundamental frequency (fig. 2, lower plot). If a sub-multiple of a real fundamental is chosen, the overtone series of such a false f_0 exhibits many missing partials, and the average power decreases rapidly. If an octave error takes place, i.e. the candidate is a multiple of a real fundamental, the overtone series is virtually complete, however many real partials are omitted. Since these odd partials have

usually high energy in natural sounds, the resulting value of MOP is also lower. Many experiments show that this very simple technique is very successful at detection of the prominent fundamental frequency and is quite immune to octave errors.
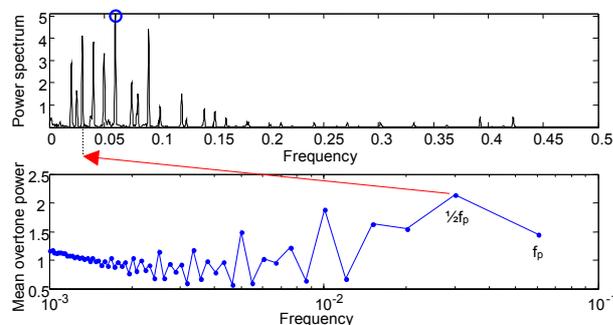

Figure 2. Fundamental frequency estimation by the detection of maximum mean overtone power (MOP)

Estimation of multiple fundamental frequencies is a result of an iterative process, wherein subsequent f_0 estimations are done after removing of the already resolved harmonic series from the signal spectrum. This removal is based on a more reliable and exact values of f_0 and β obtained from the harmonic model, after the existence of the harmonic series has been confirmed by successful tracking. If tracking is unsuccessful, it means that the most prominent peak which initiated the whole process is an unreliable source of information and should be neglected. In our implementation, a soft damping function is applied to reduce the power of this peak, and the f_0 estimation procedure is repeated.

Removing of already estimated harmonic series from the signal is performed in the power spectrum domain, by application of a simple harmonic filter, which is designed according to (5),

$$H_{f_0,\beta}(f) = \prod_{n=1}^{N} \left[1 - \lambda_n \exp\left(-\frac{(f - f_n)^2}{2\sigma^2} \right) \right] \qquad (5)$$

Chances are that also peaks from other harmonic series are reduced by multiplication with $H(f)$, if they coincide with overtones of the already existing ones. In order to avoid this situation, the principle of spectral smoothness [7] may be employed to determine the optimal weights λ_n of the individual bands.

3. ITERATIVE ESTIMATION OF HARMONIC SERIES

Proper resolving of the harmonic series from given signal spectrum involves estimation of the true fundamental frequency, f_0, as well as the inharmonicity factor β that govern the frequency relations of harmonic partials according to the model (2). This difficult task in general [5] is much easier having a preliminary estimate of the fundamental frequency. As the pre-estimated f_0 is only an approximate and β is unknown, we employ an iterative search procedure that is inspired by the EM (expectation maximization) method. The initial value of β is assumed to be 0. Alternatively, it is possible to examine the values of β estimated in previous frames. Each iteration of the adjustment algorithm consists of:

- calculation of the frequency limits for searching the overtones of the current f_0 estimate,

$$f_n^{\text{min}} = f_n(1 - \varepsilon), \quad f_n^{\text{max}} = f_n(1 + \varepsilon) \qquad (6)$$

- detection of all sinusoidal partials in each range of $f^{\text{min}} \ldots f^{\text{max}}$,

- exact estimation of the partial frequencies [10]

- finding a partial that most closely matches the appropriate value of f_n in each range,

- updating the estimates according to

$$\hat{f}_0 = E\left\{ \frac{f_n}{n \sqrt{1 + \hat{\beta}\,(n^2 - 1)}} \right\}, \quad \hat{\beta} = E\left\{ \frac{f_n^2 - n^2 \hat{f}_0^2}{n^2 - 1} \right\} \quad (7)$$

Our experiments show that in most cases a convergence of the above adjustment procedure is achieved in 2-3 iterations, depending on the harmonic density of the audio spectrum and the value of ε, usually in the range of 10^{-3}.

4. PREDICTION-BASED TRACKING OF HARMONIC PARTIAL GROUPS

Several tracking algorithms for the sinusoidal model have been proposed hitherto. Lagrange et al [6] introduced a tracking technique that exploits the principles of linear prediction (LP) of speech for the prediction of partial frequencies. In fact, these

frequencies are strictly related to pitch, and pitch changes are often governed by simple dynamics of player's gestures in natural as well as many electronic sounds. Lagrange et al successfully applied the Burg variant of the LP predictor for tracking individual partial frequencies changing in time. The idea for the trajectory continuation rule is to select such data points from the sinusoidal analysis that in the current frame are closest to the predicted values on each of the already known trajectories in the previous frames.

For tracking of harmonically-related groups of partials we propose a vector generalization of the LP-based tracker. In the ideal case, the trajectories of individual partial frequencies are scaled versions of the trajectory of f_0. Therefore, it is possible to make the overall prediction-based tracking more robust by employing a prediction error measure that is minimized globally for a group of partials. In our implementation, the Burg algorithm is modified in such a way, that the error is calculated as a vector inner product. At each iteration, the reflection coefficient r_k [6] is common for all partial predictors, and expressed as

$$ r_k = \frac{-2 \sum_{n=k}^{N-1} \mathbf{e}_{k-1}^f(n) \cdot \mathbf{e}_{k-1}^b(n-1)}{\sum_{n=k}^{N-1} \left\| \mathbf{e}_{k-1}^f(n) \right\|^2 + \left\| \mathbf{e}_{k-1}^b(n-1) \right\|^2}, \quad (8) $$

where for each sample n, $\mathbf{e}_{k-1}^f(n)$ is a vector of k-th order forward prediction errors of all partials within a harmonic group, and $\mathbf{e}_{k-1}^b(n)$ is a corresponding vector of backward prediction errors.

The important issue of every partial tracking algorithm is how to cope with singularities, such as incomplete or obviously erroneous data (missing partial information), beginnings and ends of trajectories, crossing trajectories, etc. Many of these are efficiently resolved by the introduction of zombie states of sinusoidal trajectories. Sometimes a chain of consecutive zombies needs to be inserted in order to maintain a continuity of a group of trajectories. In the case of LP-based tracking, the values of zombie states may be calculated by feeding the predictor with a sequence of zeros, however such approach quickly yields a sequence of extrapolated data that is heavily biased (fig. 3). By employing vector tracking, the calculation of missing data is supported by the tracking results of other partials in the same harmonic group.
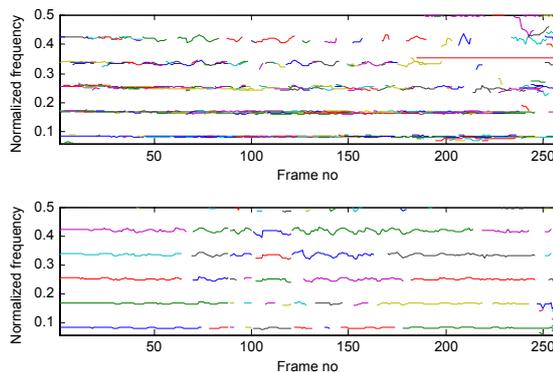

Figure 3. Comparison of unconstrained tracking of individual sinusoidal partials (above) and joint tracking of harmonic group (below)

5. EXTENSIONS OF THE HARMONIC MODEL

5.1. Complex amplitude envelope

The idea of the complex envelope comes from the observation that certain residual frequency error of each sinusoidal partial may be represented by a low-frequency oscillatory component. The extension of the harmonic model proposed here is to replace the traditional low order piece-wise approximation of partial envelopes $A_{n,k}(t)$ in (2) with narrowband complex signals. These are obtained from the original signal by single side band demodulation using a continuously changing frequency, following the frequency trajectory of each partial (9). The demodulation product should be lowpass-filtered in order to remove the unwanted side bands related to the remaining portions of signal spectrum

$$ A_{n,k}(t) = \left[x(t) \exp\left(-j\, 2\pi \int_0^t f_{n,k}(\tau)\, d\tau \right) \right] * h_{LP}(t). \quad (9) $$

The bandwidth of the filter $h_{LP}(t)$ is not critical, however it should be sufficiently narrow to effectively separate a single sinusoidal partial from other partials.

This demodulation results in straightening of the variable spectral content surrounding each harmonic partial, so that its instantaneous frequency is close to zero. The magnitude of the complex envelope signal corresponds to the real envelope of the particular partial,

while the phase is related to the frequency estimation error as well as any deviations of the real frequency from the smooth low-order curve used within the model. Such narrowband signals may be very efficiently represented, e.g. using transform coding [11]. They are also very easily transformable for any kinds of signal manipulations, like time stretching, pitch shifting etc.

Experimental results show that such an extension of the harmonic model yields very naturally sounding reconstructed signals, thus it is possible to use this extended model in a broad range of applications, including high quality audio data compression [11].

5.2.   Individual frequency residual

The formulation of the traditional harmonic model (1) assumes that all partial frequencies are strictly related to the fundamental frequency. Synthesis of an audio signal from such a model often results in artificially sounding signals due to rigid harmonic relations of overtones. In real world musical sounds, certain frequency fluctuations may be observed due to various physical phenomena of vibrating objects, like stiffness, residual modulation, coupling of oscillation modes, etc. In order to cover these natural sound properties in our model, we propose to extend the formula of the polyphonic harmonic model (2) in such a way, that each partial is allowed an individual frequency deviation from the harmonic law (10).

$$f_{n,k}(t) = n \, f_k(t) \, \sqrt{1 + \beta_k \, (n^2 - 1)} + \vartheta_{n,k}(t) \qquad (10)$$

The frequency deviation (residual) component, $\vartheta_{n,k}(t)$ is a band-limited and value-bounded random process incorporating the above mentioned irregularities as well as the frequency estimation errors.

One interesting way to obtain the values of the residual components $\vartheta_{n,k}(t)$ is to extract the complex amplitude envelopes $A_{n,k}(t)$, and perform an instantaneous frequency analysis. Since the envelopes are narrowband and (presumably) mono-component, the instantaneous frequency is simply a derivative of their phase [12]. In certain situations, transients and excessive noise may cause the derivative to become instable. In our implementation we use a combination of median filtering, clamping and linear smoothing to keep the estimated $\vartheta_{n,k}(t)$ well-behaved.

The values of $\vartheta_{n,k}$ should be stored together with other model data for all applications requiring high quality reconstruction. For compression purposes it is possible to either apply a lossy waveform coding to the $\vartheta_{n,k}(t)$, or to apply a parametric coding principle, i.e. represent only the basic statistical properties of the frequency residual and employ a random generator at the decoder side to rebuild a similar signal.

6.   EXPERIMENTAL RESULTS

The polyphonic harmonic model described in this paper has been implemented in software in the form of a Matlab toolkit. Such implementation allowed for throughout testing of the behavior of its various components in many different conditions. For example, the model has been applied for analysis and coding of polyphonic audio at the range of bit rates of 16-32 kb/s (fig. 4). We used the EBU SQAM reference CD [13] as a source of many testing excerpts.
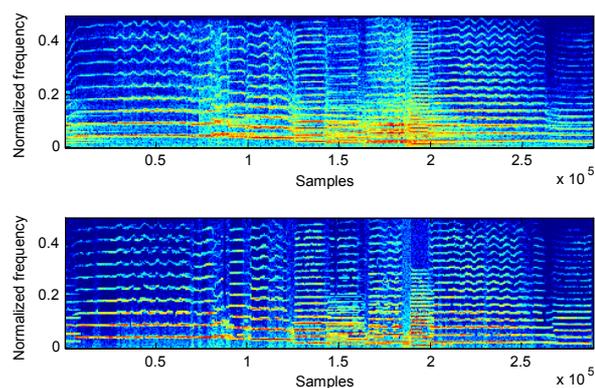

Figure 4. The spectrograms of the original music excerpt (above) and the reconstructed signal from the sinusoidal model encoded at 29kb/s (below)

Depending on the complexity of audio source material the resulting quality of reconstructed music samples was moderate to very good, showing a significant advantage in terms of compression efficiency over the non-harmonic sinusoidal model.

Some informal listening test were also conducted in order to compare the degree of quality improvement offered by the two extensions proposed in section 5. The common observation is that replacing of a real-valued piece-wise linear envelopes by complex narrowband envelopes with a bandwidth of about 5 Hz results in a tremendous advantage in the reconstructed audio

fidelity. Using linear envelopes and encoding the frequency residual also resulted in a high quality reconstruction. The perceptual difference between these two variants was minor and often hard to notice.

It may be admitted that the frequency residual is just a simplified representation of the complex envelopes, wherein the magnitude is approximated by a piece-wise linear function. Our listening test results show that minor phase discrepancies are more important for the human auditory system than amplitude discrepancies of individual harmonic partials.

7.   CONCLUSIONS

A simple implementation of a polyphonic harmonic model for music analysis has been presented in this paper. A combination of a bootstrap approach involving multiple f_0 pre-estimation and iterative adjustment of the harmonic series parameters proves to be an efficient modeling tool. Two extensions of the harmonic model have been also described. Informal listening tests show a significant quality improvement may be achieved by the introduction of complex amplitude envelopes or individual frequency residual functions to the model data.

8.   ACKNOWLEDGEMENTS

9.   REFERENCES

[1] X. Serra, *A system for sound analysis/trans-formation/synthesis based on a deterministic plus stochastic decomposition*, Ph.D. Thesis, Stanford University, Stanford,1989

[2] Vincent, E.; Plumbley, M.D., "A prototype system for object coding of musical audio", *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* pp. 239-242, 16-19, Oct. 2005

[3] Gribonval, R.; Bacry, E., "Harmonic decomposition of audio signals with matching pursuit", *IEEE Transactions on Signal Processing*, vol.51, no.1, pp. 101-111, Jan. 2003

[4] Krstulovic, S.; Gribonval, R.; Leveau, P.; Daudet, L., "A comparison of two extensions of the matching pursuit algorithm for the harmonic decomposition of sounds", *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* pp. 239-242, 16-19, Oct. 2005

[5] Wen Xue, Mark Sandler, "Sinusoid modeling in a harmonic context", *Proc. 10th Int. Conference on Digital Audio Effects, DAFx-07*, Bordeaux, 2007

[6] M. Lagrange, S. Marchand, M. Raspaud, J-B. Rault, "Enhanced Partial Tracking Using Linear Prediction", *Proc. 6th Int. Conf. Digital Audio Effects DAFx-03*, London, UK, 2003

[7] A.P.Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness", *IEEE Transactions on Speech and Audio Processing*, vol.11, no.6, pp. 804-816, Nov. 2003

[8] R. Meddis, M.J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification", *Journal of the Acoustical Society of America*, vol. 89, no. 6, June 1991.

[9] P.J. Walmsley, S.J. Godsill, P.J.W. Rayner, "Bayesian modelling of harmonic signals for polyphonic music tracking", *Cambridge Music Processing Colloquium*, Sept. 1999

[10] F. Keiler, S. Marchand, "Survey on extraction of sinusoids in stationary sounds", *Proc. 5th Int. Conf. on Digital Audio Effects, DAFx02*, Hamburg, Sept. 2002

[11] M. Bartkowiak, "A complex envelope sinusoidal model for audio coding", *Proc. of the 10th Int. Conference on Digital Audio Effects, DAFx-07*, Bordeaux, 2007

[12] B. Boashash, "Interpreting and estimating the instantaneous frequency of a signal—Part II: algorithms", Proceedings of the IEEE, vol. 80, pp. 539–569, April 1992

[13] .European Broadcast Union, "SQAM, Sound quality assessment material recordings for subjective tests", CD and Technical Report No. 3253-E, Brussels, Belgium