Periodicity and missing feature theory in audition

Alain de Cheveigné (CNRS/ATR-HIP)

Abstract

Sounds that are generated by exciting a resonator usually have a timbre that depends on the characteristics of the resonator (transfer function). In the case of vowels, the resonator is the vocal tract, in the case of musical instruments it may be the resonance tube (woodwinds, etc.) or the body of the instrument (violin, etc.). For a constant resonator, timbre is relatively independant from source characteristics such as fundamental frequency. However, it is not evident how resonance characteristics can be extracted from the waveform. Short-term spectra are strongly affected by the fundamental periodicity, and the same is true of auditory representations such as basilar-membrane excitation pattern. This talk is divided into two parts. The first is a sort of "tutorial" on spectral estimation, the problems posed by harmonicity, the notion of pitch-dependent smoothing and its limits, and the idea of timbre pattern-matching using missing-feature techniques. The second part describes a vowel-perception model based on these ideas.

1 Harmonicity and spectral envelope

1.1 The aim

Take a sound produced by exciting a resonator with a periodic source. It could be for example a vowel, produced by exciting the vocal tract by a train of glottal pulses. It is our common experience that the timbre (for example the vowel quality) depends upon the *spectral envelope* of the sound, defined as a function of frequency $H(\omega)$ that determines the amplitude at each partial frequency. Timbre hardly depends on phase. It also changes rather little with changes in the fundamental frequency (F_0) of excitation, that strongly affect the pitch. Conversely, timbre varies widely with changes in the spectral envelope that hardly affect the pitch. To a first approximation, pitch and timbre are independent percepts, function respectively of the F_0 and the magnitude spectral envelope. The spectral envelope of the waveform is that of the source multiplied by the *magnitude transfer function* of the resonator. To the extent that the spectral envelope of the source is constant (or better still, flat) the timbre of a sound reflects the resonator that produced it. This may give timbre its ecological value: it tells us about the nature (shape, size) of the object that produced the sound.

Our aim is to estimate this spectral envelope from the waveform. The motivation is two-fold. First, to obtain a procedure to "measure the timbre" of a waveform, for classification purposes or as a model of timbre perception. Second, to allow resynthesis after manipulation of source, timbre, or duration characteristics (sound morphing).

The definition of a "spectral envelope" (a function of frequency that determines the amplitude of each partial) is clear in the context of production. It does not follow that it can easily be estimated from the waveform. Indeed, this talk is motivated by the various difficulties that are involved. Given those difficulties, it may be useful to give "spectral envelope" the wider definition of a function of frequency that (a) is independent of F_0, and (b) "reflects" the resonant characteristics in some useful fashion. Obviously, if property (a) is not granted, the correspondence between envelope and resonator is F_0-dependent, and property (b) cannot be insured. Nevertheless, (a) is often further relaxed to mean "no visible harmonic structure", in other words simply that the spectral envelope is "smooth". Whether this very loose definition is acceptable depends on the application.

1.2 Issues

In this section, a number of issues involved in spectral envelope estimation are examined one by one. To simplify the discussion, the source (periodic or otherwise) is supposed to have a flat spectral envelope. The spectral envelope of the waveform $S(f)$ is thus identical to the magnitude transfer function $H(f)$ of the resonator. The two terms are used interchangeably.

1.2.1 An intuitive argument

A fixed resonator is characterized by a one-dimensional function of frequency, $H(f)$. This is commensurable to the one-dimensional waveform $s(t)$, and one can expect some success in estimating the former from the latter, within limits that will soon be made clear.

Suppose instead that the resonator is time-variant. Strictly speaking, a "transfer function" is meaningful only for a linear time-invariant (LTI) system. Suppose nevertheless that it can be extended to the time-variant case. The transfer function $H_t(f)$ is now a two-dimensional function of time and frequency. Intuitively, we can expect to have difficulty in estimating all details of the two-dimensional function $H_t(f)$ from the one-dimensional signal $s(t)$.

1.2.2 The sampling theorem

Much use will be made of the *sampling theorem* applied to spectral envelopes. In its standard form, the sampling theorem states that a signal $s(t)$ sampled at intervals of $T_s = 1/f_s$ can be reconstructed perfectly from its samples if it is band-limited to less than the Nyquist frequency, $f_s/2$.

The same theorem can be applied to spectral envelopes, replacing time t by frequency f, and frequency by time, or more appropriately *lag* (also known as *quefrency*). The theorem says that a spectral envelope sampled at intervals of $f_0 = 1/T_0$ can be reconstructed from the samples if it is band-limited to lags shorter than the "Nyquist lag" $T_0/2$, that is, if the Fourier Transform of the spectral envelope is zero beyond that lag.

1.2.3 Envelope shape and impulse response

It is equivalent to know the spectral envelope itself or any invertible function (square, log, etc.) of it. One can go from one to the other by applying the function or its inverse. In particular, the magnitude spectral envelope is entirely specified by the *squared magnitude transfer function*, the Fourier Transform of which is equal to the *autocorrelation function of the impulse response* (IRACF) of the filter. If the magnitude spectral envelope is sampled at intervals of $f_0 = 1/T_0$, and if the IRACF is zero beyond the Nyquist lag $T_0/2$, the squared envelope (and therefore the envelope itself) can be reconstructed from the samples.

A possible source of confusion should be pointed out. The various transforms of the envelope (square, log, etc) have different compositions in the lag-domain. It may happen that one is band-limited while the others are not. The sampling condition should thus be restated: the envelope sampled at f_0 can be reconstructed *if there exists a known invertible function* of the envelope that is band-limited. It may seem confusing that different functions of the spectral envelope (magnitude, power, log) have different lag-domain contents, but the same is of course true of the spectra of time-domain functions. As an example, the third power of a sine-wave (band-limited in the frequency domain) contains an infinite set of harmonics, products of non-linear distortion.

In particular, it is worth noting that some function of the spectral envelope might be band-limited, whereas the squared magnitude envelope might not. It follows that, while the condition that the IRACF is zero beyond $T_0/2$ is *sufficient* for reconstruction from samples spaced at intervals of $f_0 = 1/T_0$, it is not a *necessary* condition.

1.2.4 Time-invariant resonator

Supposing that the resonator is constant, let us consider three possible excitation functions: a single pulse, white noise, and a periodic pulse train.

If the resonator is excited by a single pulse at the origin of time, the waveform $s(t)$ is simply the impulse response $h(t)$. If the waveform can be observed for a duration longer than the support of $h(t)$, the spectral envelope can be perfectly estimated by calculating the Fourier Transform of the IRACF and then taking its square root.

If the resonator is excited by a pulse train of fundamental frequency $f_0 = 1/T_0$, its spectral envelope $H(f)$ is *sampled* at intervals of f_0. From the sampling theorem we know that it is perfectly represented by the samples if it is band-limited to lags shorter than the Nyquist lag $T_0/2$. A sufficient (but not necessary) condition is that its impulse response be shorter than $T_0/2$, so that the IRACF is zero beyond $T_0/2$. In the general case, however, the spectral envelope is *not* band-limited. The spectral envelope is incompletely represented by the samples and cannot be reconstructed perfectly. More on this problem later on.

If the resonator is excited by white noise, the spectral envelope is the product of the transfer function by a spectrum that is flat. The spectral envelope can be calculated directly by taking the Fourier Transform of the waveform over infinite time. In practice however, any estimate based on a window with finite length is noisy, with an amount of noise that depends on the window length.

Clearly, the "best" excitation is a single pulse. Periodic and white-noise excitation have complementary advantages and disadvantages. Periodic excitation samples the transfer function at discrete points. The sampling is sparse, but each sample is accurate. White noise on the other hand "samples every point", but with a degree of uncertainty that depends on the temporal averageing involved in the calculation.

1.2.5 Time-variant resonator

If the resonator varies in time (as is the case of the vocal tract in speech), the problem is more difficult. Parameters of the resonator vary in time, and it is tempting to conclude that its transfer function also varies in time. However, strictly speaking, the notion of "transfer function" is only defined for linear time-invariant (LTI) systems, for which complex exponentials are eigen-vectors. A time-varying system breaks a fundamental rule. There are several ways out of this predicament. (1) Assume that variations are slow relative to the maximum duration of the impulse response, and that errors are likely to be small. (2) Try to estimate a set of time-varying parameters of the resonator, based on a model, rather than the transfer function. (3) Try to estimate a "time-varying spectral envelope" such that, given a particular synthesis technique, the signal can be accurately synthesized based on this time-varying envelope. The drawback of the latter two approaches is that they are dependent on particular production or synthesis models.

Let us suppose that we can somehow define a time-varying envelope $H_t(f)$. It can be visualized as a surface in 3-D space, function of frequency and time. If it is based on assumption (1), variations along the time axis must be smooth and slow. If it is based on assumption (3), this assumption is not necessary, supposing that the synthesis technique can handle a quickly-varying envelope.

It makes little sense to excite such a resonator with a single impulse. Even if the resonator varies slowly enough for the impulse response to be meaningful, it only captures the shape at one point in time. Excitation with noise makes more sense, but the amount of time available for averageing limits estimation accuracy.

In the case of a periodic pulse-train excitation, one can conceive of the function $H_t(f)$ as being sampled in both time and frequency. As it were, the surface in 3-D space is sampled at discrete points regularly spaced on both axes. The limits of this image should be pointed out immediately: perfect resolution along the frequency axis depends on stationarity along the time axis: it can be obtained only if $H_t(f)$ is constant in time (time-invariant filter). Conversely, perfect resolution along the time axis can be obtained only if the impulse response has zero length, that is, if $H_t(f)$ is constant in frequency (wide-band attenuator). These two conditions are mutually exclusive. In the general case of a time-varying resonator, the 2D sampling grid is necessarily "fuzzy".

Suppose nevertheless that the surface can be sampled precisely, at intervals of T_0 along the time axis, and $f_0 = 1/T_0$ along the frequency axis. The surface can be reconstructed if it is bandlimited on both axes: to lags shorter than $T_0/2$ for the shape along the frequency axis, and to frequencies smaller than $f_0/2$ along the time axis. [A more rigorous discussion should involve a 2-D sampling theorem].

In the general case, chances are that $H_t(f)$ is *not* band-limited along either axis. In that cas it is incompletely represented by the samples and cannot be reconstructed perfectly. It is worth noting however that the samples themselves, although sparse, are accurate. This is in contrast with the case of white-noise excitation, for which all data are noisy.

The undersampling problem is most severe in situations such as speech, where the f_0 varies over a wide range: $H_t(f)$ is likely to be undersampled along the frequency axis when f_0 is high, and along the time axis when f_0 is low.

1.3 Solutions

This section reviews several approaches to the problem of envelope estimation. The choice of approach depends on two factors: (b) whether the shape of the resonator's transfer function $H_t(f)$ is sufficiently smooth (band-limited) to allow adequate sampling, and (b) the application: morphing *vs* classification.

1.3.1 Harmonic excitation: two cases

Suppose the resonator is excited by a periodic pulse train of fundamental frequency f_0. One can distinguish two cases. In the first case, the spectral envelope $H_t(f)$ is bandlimited to less than the Nyquist lag $T_0/2$ along the frequency axis, and less than the Nyquist frequency $f_0/2$ along the time axis. In this case, the spectral envelope can be perfectly reconstructed, *if the train of samples is filtered perfectly to remove out-of band components*. The smoothing task, which is not trivial, is the object of pitch-period smoothing techniques to be described soon.

The second case is when $H_t(f)$ is not band-limited along the time and/or frequency axis. In this case it *cannot* be reconstructed from the samples. Accurate estimation of the spectral envelope cannot be performed in this case. Depending upon the application, there are two reasonable courses that may be taken.

The first is to apply pitch-period smoothing techniques to remove out-of-band components (that are certainly incorrect). The result nevertheless differs from the true envelope, because aliasing causes out-of-band components of the original envelope to be folded back into the band, where they mix with genuine in-band components. This course is reasonable if the application is for example morphing.

A second course may be more reasonable if the application involves pattern matching (classification). Because of aliasing, smoothing the samples produces an envelope that is complete but incorrect. On the other hand, recall that the samples themselves are sparse but accurate. It makes sense, therefore, to avoid smoothing and perform pattern matching directly on the samples, restricting the match to the samples themselves. This is an example of *missing feature* techniques. Missing feature techniques do not prevent the reliability of pattern matching may be affected by the sparse sampling: patterns that differ at points other than the samples cannot be discriminated, and matching is overall more sensitive to noise. Nevertheless, missing-feature techniques allow the systematic errors that arise from aliasing to be avoided.

1.3.2 Low-pass filtering for reconstruction

Suppose that we are in the first case: the resonator is excited by a periodic pulse train with a fundamental frequency f_0, and the envelope $H_t(f)$ is appropriately band limited. The envelope can be reconstructed from the samples by low-pass filtering them in both time and frequency to remove out-of-band components. For reconstruction to be accurate the filtering must be perfect: flat transmission up to the Nyquist frequency (or lag), infinite rejection beyond.

However, actual filtering is never perfect. Imperfect filtering may cause two problems, of unequal severity: (1) The pass-band is not flat, so the reconstructed envelope is a low-pass filtered version of the original envelope. (2) Out-of-band components are not perfectly attenuated.

The first problem is relatively minor. The second is more serious, mainly because it implies that *the reconstructed envelope is wide-band*, even if the original envelope is narrow-band. Even if the original envelope is band-limited, the reconstructed envelope may contain out-of-band components due to imperfect low-pass filtering. This problem is troublesome particularly if the the estimated envelope must be *re-sampled*. There are two typical situations where such is the case: *down-sampling* (to a "frame-rate" along the time axis, or to frequency bands along the frequency axis), and *resynthesis* after manipulation of F_0, duration or spectral envelope. Distortion caused by aliasing degrades the quality of resynthesis.

A certain degree of low-pass filtering is inherent in the Fourier transform. Integration over a time-window implies low-pass filtering along the time axis. Conversely, the limited spectral resolution implies a form of low-pass filtering along the frequency axis.

However this low-pass filtering is not really sufficient: there is typically too much resolution along both axes.

1.3.3 Pitch-period smoothing in the time domain

Let us back up a bit, and consider what happens if we apply a particular analysis at every point in time (running analysis). Take the case of a coefficient $c(t)$ produced by the analysis at time t. The analysis could be the Fourier Transform of the previous paragraph [$c(t)$ being a spectral coefficient corresponding to a given frequency], or it could be simpler (energy) or more complex (cepstrum, LPC, etc.). In the real world, such analyses are usually performed at a reduced rate (frame rate) but that is not necessary for their principle, nor desirable for our purpose. Instead we assume *running* analyses indexed by time (what that means in practice is that the analysis is repeated at intervals of one period of the waveform sampling rate).

Assume that the analysis that produces $c(t)$ is supposed to tell us about the macroscopic properties of the signal. Ideally, $c(t)$ should be constant if $s(t)$ is stationary. In practice that may not be the case when smoothing is insufficient, but there is one thing we can guarantee. If the signal $s(t)$ is periodic with period T_0, and the analysis is deterministic and time-invariant, $c(t)$ is also periodic with period T_0 (in a wide sense, including the cases where $c(t)$ is constant or periodic with a period that is a fraction of T_0).

Fortunately, there exists a filter that is useful for this situation. Its impulse response, $h(t) = U(t) - U(t - T_0)$ is shaped like a square window of duration T_0. Its transfer function is shaped like a sinus cardinal with zeros at all multiples of $f_0 = 1/T_0$ except 0. Applied to a periodic signal such as $c(t)$, the filter removes all harmonics of the fundamental and leaves only the zero-hertz component, as is desired. This property extends usefully to the convolution of this window to any other. For example a convolution of the square window with itself produces the *Bartlett* window, which has a similar desirable property.

Such filtering is referred to here as "pitch-period smoothing in the time-domain". To apply it requires an estimate of the fundamental period.

1.3.4 Pitch-period smoothing in the frequency domain

A similar form of smoothing may be applied in the frequency domain. The spectrum is convolved by a square window of width f_0 (or a window derived from it by convolution). This removes all lag-domain components multiple of the period T_0. Once again, an estimate of the fundamental period is required.

1.3.5 Time-frequency smoothing: STRAIGHT

Kawahara's STRAIGHT system implements the previous ideas with several refinements. Smoothing is performed simultaneously in time and frequency domains. The projection of the smoothing kernel along both axes is shaped like a Bartlett window (rather than square). Smoothing is performed on an arbitrary invertible non-linear transform of the magnitude spectrum (to allow simulation of smoothing of loudness, etc.). The system includes various "tricks" to enhance processing. It also comes complete with an F_0-estimation module (TEMPO) and a module for the fine control of source properties for resynthesis (SPIKES).

1.3.6 Caveats of pitch-period smoothing

For a *stationary* periodic excitation and resonator, pitch-period smoothing does much better than any other form of smoothing. Out-of-band harmonic components are perfectly eliminated with a minimum loss of resolution in time or frequency (especially if the shortest possible windows - square - are used). However our analysis, accurate for the stationary case, may not extend perfectly to a *time-varying* resonator or fundamental frequency. In practice, we can expect the quality of smoothing to depend on how close we are to the stationary case.

Pitch-period smoothing reconstructs the envelope perfectly if it is appropriately bandlimited. Along the frequency axis, the Nyquist lag is $T_0/2$, for the time axis the Nyquist frequency is $f_0/2$. Since $T_0 = 1/f_0$, these two limits are of course linked. If f_0 varies over a wide range (as in speech), the chances are large that one or other limit is crossed at times.

When the envelope is not appropriately bandlimited, the reconstructed envelope necessarily differs from the original envelope, because of aliasing. The error is f_0-dependent, in magnitude and also in shape. Depending on the application, the result may nevertheless be acceptable. For the purpose of resampling, for example, the pitch-period-smoothed envelope has the desirable feature that it is band-limited (see above). For the purpose of pattern-matching, on the other hand, it may be better to avoid smoothing and apply missing-feature techniques, to be discussed soon.

1.3.7 Practical considerations

Computational cost Typical spectral analysis is based on FFTs performed at a certain frame rate, so that successive FFT windows overlap in time. If N is the FFT window size, and K is the overlap factor (meaning that the frame period is N/K), then the cost per sample is on the order of

$$Klog(N) \tag{1}$$

(assuming that the cost of an N-point FFT is $Nlog(N)$). The factor K is usually small. Pitch-period smoothing, to be effective, must be performed *before* down-sampling to the frame rate. Spectral analysis must therefore be performed at every sample, at a cost of:

$$Nlog(N) \tag{2}$$

Pitch-period smoothing is thus potentially rather expensive. One way to reduce cost is to perform a running DFT (Rabiner and Schaffer, 1978). The cost is on the order of

$$N \tag{3}$$

which is still more expensive than performing FFTs at a reduced frame rate as in classic analysis. An advantage of the running DFT is that N does not need to be a power of 2. A drawback is that the analysis window must be square (or a convolution of square windows).

F_0 estimation F_0 estimation is critical to pitch-period smoothing. Reliable F_0 estimation is known to be difficult for speech, an important application field. However the task may be actually easier than it seems.

Estimation errors can be divided into three types. (1) Random errors, that occur when periodicity is poor. (2) Subharmonic errors that often occur despite very clean periodicity, as the result of small sampling errors or localized noise. (3) Harmonic errors, due to the dominance of a single strong harmonic.

Errors of type (1) are not serious, in the sense that little is to be gained by pitch-period smoothing when the signal is not periodic. The main problem is that random switching from one estimate to another introduces "noise" in the estimation process. A solution might be to set the estimate to a fixed "default" value when a periodicity measure falls below a certain threshold. Remain the problems of choosing the best default, and handling the transition to and from this value each time the threshold is crossed.

Errors of type (2) are serious for smoothing along the frequency axis, as they produce a window that is too small. They are of little consequence for smoothing along the time axis, as a window that is two or three times the period is effective (the only penalty is oversmoothing). Errors of type (2) can be virtually eliminated by "biasing" the F_0-estimation algorithm towards short estimates.

Likewise, errors of type (3) are serious for smoothing along the time axis, but not the frequency axis. They can be eliminated by biasing the F_0-estimation algorithm towards long estimates.

The key to reliable smoothing is thus to use two F_0 estimates, each with a different bias, for temporal and spectral smoothing respectively.

1.3.8 Severe undersampling: what to do?

Consider now the case where the envelope $H_t(f)$ is *not* adequately bandlimited. Whatever the smoothing, the estimated envelope is distorted due to aliasing, as out-of-band components are folded and summed together with genuine in-band components. Depending on the application, two approaches may be of use. *Missing feature theory* is appropriate for classification, whereas *model fitting* may provide a solution when a smooth envelope is required for resynthesis.

Missing feature theory Suppose that the task is to classify spectral envelopes by pattern-matching with pre-determined templates. The available samples may be matched directly to the templates, using a non-uniform weighting function that restricts the calculation of the distance function to the samples, with zero weight applied to all other points. As long as smoothing or interpolation are *not* performed, this method avoids distortion due to aliasing. Assuming that the samples themselves are accurate, a perfect match can be made with the corresponding points of the appropriate template. One can argue that this is the best approach to classification, and that smoothing, interpolation or model-fitting can provide no improvement. Available information is entirely contained in the samples: interpolation can create no new information. Interpolation can be construed as an *informed guess* of what missing data points should look like. If that guess is wrong, the answer may be incorrect.

Model fitting The "Nyquist barrier" can be broken if an underlying model of the envelope is available (as long as the sample points are sufficiently numerous and accurate to constrain the model). An example might be an articulatory model of the vocal tract that constrains the envelope to one that can be produced physiologically. The constraints might apply to the spectral shape, based on the range of possible vocal tract shapes, and also to its variation in time based on constraints on how the articulators can move.

Following the philosophy of missing feature theory, model fitting should be restricted to available sample points. However these do not necessarily have to be equally spaced.

The hypothesis of harmonic excitation, while it gives us a convenient way to determine the position of sample points, is not necessary.

The drawback of missing feature and model approaches is that they require underlying models or templates. They are of no help to produce a universal tool for envelope estimation.

2 Missing feature model of vowel perception

The missing feature model of vowel perception builds on the previous analysis. The model is described in detail in a draft that can be downloaded from the following address: $http://llf.linguist.jussieu.fr/ alain/ps/miss_v ow.ps.gz$

Abstract Vowel identity correlates well with the shape of the transfer function of the vocal tract, in particular the position of the first two or three formant peaks. However in voiced speech the transfer function is *sampled* at multiples of the fundamental frequency (F_0), and the short-term spectrum contains peaks at those frequencies, rather than at formants. It is not clear how the auditory system estimates the original spectral envelope from the vowel waveform. Cochlear excitation patterns, for example, resolve harmonics in the low frequency region and their shape varies strongly with F_0. The problem cannot be cured by smoothing: lag-domain components of the spectral envelope are aliased and cause F_0-dependent distortion. The problem is severe at high F_0s where the spectral envelope is severely undersampled. This paper treats vowel identification as a process of pattern recognition with *missing data*. Matching is restricted to available data, and missing data are ignored using an F_0-dependent weighting function that emphasizes regions near harmonics. The model is presented in two versions: a frequency-domain version based on short-term spectra, or tonotopic excitation patterns, and a time-domain version based on autocorrelation functions. It accounts for the relative F_0-independency observed in vowel identification.

References

[1] Cooke, M., Green, P., Anderson, C., and Abberley, D. (1994), "Recognition of occluded speech by hidden markov models," University of Sheffield Department of Computer Science technical report, TR-94-05-01.

[2] Cooke, M., Morris, A., and Green, P. (1997). "Missing data techniques for robust speech recognition.", Proc. ICASSP, 863-866.

[3] de Cheveigné, A., and Kawahara, H. (1998), "A model of vowel perception based on missing feature theory," ATR-HIP technical report, TR-H-252.

[4] de Cheveigné, A., and Kawahara, H. (XXXX). "Missing feature model of vowel perception," JASA (submitted)

[5] de Cheveigné, A. (1998). "The auditory system as a separation machine.", Proc. ATR workshop on events and auditory temporal structure, 1-7.

[6] Kawahara, H. (1997). "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited.", Proc. ICASSP, 1303-1306.

[7] Lippmann, R. P., and Carlson, B. A. (1997). "Using missing feature theory to actively select features for robust speechj recognition with interruptions, filtering, and noise.", Proc. ESCA Eurospeech, KN-37-40.

[8] Morris, A. C., Cooke, M. P., and Green, P. D. (1998). "Some solutions to the missing feature problem in data classification, with application to noise robust ASR.", Proc. ICASSP, 737-740.

[9] Rabiner, L. R., and Schafer, R. W. (1978). "Digital processing of speech signals," Englewood Cliffs, NJ, Prentice-Hall.