Can Temporal Fine Structure and Temporal Envelope be Considered Independently for Pitch Perception?

Nicolas Grimault

Abstract In psychoacoustics, works on pitch perception attempt to distinguish between envelope and fine structure cues that are generally viewed as independent and separated using a Hilbert transform. To empirically distinguish between envelope and fine structure cues in pitch perception experiments, a dedicated signal has been proposed. This signal is an unresolved harmonic complex tones with all harmonics shifted by the same amount of Hz. As the frequency distance between adjacent components is regular and identical than in the original harmonic complex tone, such a signal has the same envelope but a different fine structure. So, any perceptual difference between these signals is interpreted as a fine structure based percept. Here, as illustrated by very basic simulations, I suggest that this orthogonal point of view that is generally admitted could be a conceptual error. In fact, neither the fine structure nor the envelope is required to be fully encoded to explain pitch perception. Sufficient information is conveyed by the peaks in the fine structure that are located nearby a maximum of the envelope. Envelope and fine structure could then be in perpetual interaction and the pitch would be conveyed by "the fine structure under envelope". Moreover, as the temporal delay between peaks of interest is rather longer than the delay between two adjacent peaks of the fine structure, such a mechanism would be much less constrained by the phase locking limitation of the auditory system. Several data from the literature are discussed from this new conceptual point of view.

Keywords Pitch · Envelope · Fine structure

N. Grimault (✉)
Centre de Recherche en Neurosciences de Lyon, CNRS UMR 5292, Université, Lyon 1, Lyon, France
e-mail: nicolas.grimault@cnrs.fr

P. van Dijk et al. (eds.), *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing,* Advances in Experimental Medicine and Biology 894,
DOI 10.1007/978-3-319-25474-6_37

1 Introduction

Numerous works about pitch perception attempt to identify and to distinguish between relevant temporal cues as envelope and fine structure. These cues are generally viewed as independent. In fact, mathematically, these cues are orthogonal and can be extracted and separated using a Hilbert transform. Using this mathematical decomposition, some works suggest that musical pitch would rely mostly on the fine structure and that speech perception would rely mostly on envelope (Smith et al. 2002). To empirically distinguish between envelope and fine structure cues in pitch perception experiments, a dedicated signal has been proposed as early as 1956 (De Boer 1956). This signal is equivalent of an unresolved harmonic complex tone with all harmonics shifted by the same amount of Hz. As the frequency distance between adjacent components of such a signal is regular and identical than in the original harmonic complex tone, such a signal has the same envelope but a different fine structure. As a consequence, any perceptual difference between the harmonic and the shifted complex is often interpreted as a pure fine structure based percept.

Figure 1 represents the temporal information that potentially convey some pitch related temporal information. These temporal cues can be classified in three categories of periodicities. First, the periodicities of the carrier related to the delay dT1 in Fig. 1. This periodicity provides some information about the frequency of the pure tone or about the frequency of the carrier. To extract dT1 periodicities, the auditory system is required to be able to phase lock on the instantaneous phase of the signal. dT1 periodicities will be called temporal fine structure periodicities (TSF_{period}) in the following. Second, the periodicities of the envelope (delay dT2 in Fig. 1) provides some information about the F0 for harmonic complex tones. dT2 periodicities

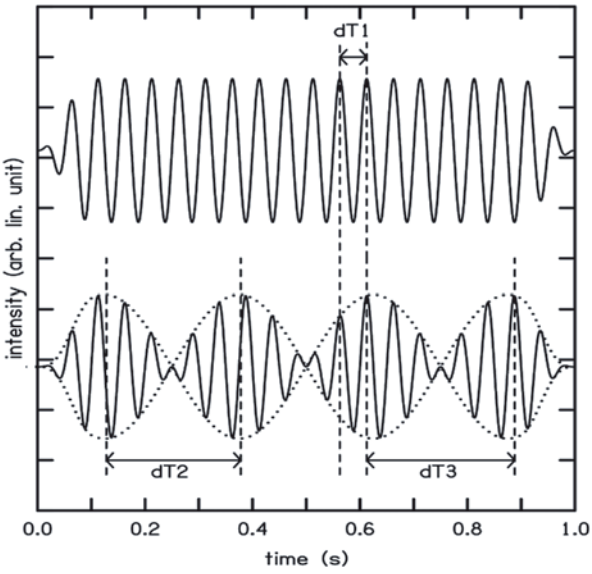

Fig. 1 Potential temporal periodicity cues usable for pith perception of a pure tone (*upper* waveform) or a shifted complex tone (*lower* waveform). dT1 is the time delay between two adjacent peaks in the temporal fine structure, dT2 is the time delay between two adjacent peaks in the temporal envelope and dT3 is the time delay between two maximums of the temporal fine structure located nearby two successive maximums of the temporal envelope. dT2 and dT3 are confounded for harmonic complex tones but different for shifted complex tones

will be called temporal envelope periodicities (TE_{period}) in the following. Third, the periodicities of the fine structure modulated by the envelope (delay dT3 in Fig. 1). This type of periodicities corresponds to the time delays between two energetic phases of the signal located nearby two successive maxima in the envelope. For harmonic complex tones, dT2 is equal to dT3 but for shifted complex tones, dT2 and dT3 are different. As dT3 is necessary a multiple of dT1, for a shifted complex, dT3 will be either equal to n × dT1 or (n + 1) × dT1 choosing the integer n to verify the chain of inequations: n × dT1 ≤ dT2 < (n + 1) × dT1. As such, dT3 depends on both dT1 (fine structure information) and dT2 (envelope information) and can then be described as a periodicity related to an interaction between fine structure and envelope. dT3 periodicities will be called interaction periodicities ($TE \times TSF_{period}$) in the following.

However, most of the time, when a harmonic complex tone and a shifted complex tone with the same envelope frequency conduct to a different pitch percept, the pitch percept is supposed to be elicited by fine structure only and to be independent of envelope. This can appear as trivial but the aim of this study is simply to convince the reader that this assumption is not true.

2 Methods

2.1 Simulation

A very basic simulation has been performed to check for the potential use of various types of periodicity cues (i.e. TSF_{period}, TE_{period} or $TE \times TSF_{period}$) with various signals used in the literature. The basic idea was to pass the stimulus through a dynamic compressive gammachirp (the default auditory filter-bank of the Auditory Image Model) (Irino and Patterson 2006) and to transform the output waveform into a spiketrain by using a threshold dependent model. This model has a higher probability to spike each time the output is over a threshold value (see Eq. 1). Moreover, a reasonable refractory period of 2500 Hz is added to the simulation. The periodicities of the spiketrain are finally extracted with an autocorrelation. So, the spiketrain generation is based on the following formula which depends on both envelope and fine structure temporal information:

$$\text{spiketrain}\,(t) = \begin{cases} 1 \text{ if } \mathrm{U}(t) \times \mathrm{TFS}_{\text{signal}>0}(t) \times E(t) \times P_{\text{refract}}(t) > \text{Thres} \\ 0 \text{ if } \mathrm{U}(t) \times \mathrm{TFS}_{\text{signal}>0}(t) \times E(t) \times P_{\text{refract}}(t) > \text{Thres} \end{cases} \quad (1)$$

Where

U(t) is an uniform intern noise between 0 and 1,

$TSF_{signal>0}(t)$ is the positive part of the fine structure at the output of an auditory filter,

E(t) is the envelop at the output of an auditory filter,

$P_{refract}(t)$ is the refractory period. This function is either equal to 0 or 1 related to a refractory period equal to 1/2500. So, i*f* $spiketrain(t_0) = 1$, $P_{refract}(t) = 0$ *for* $t_0 < t < t_0 + 1/2500$ *and* $P_{refract}(t_0 + 1/2500) = 1$,

Thres is the discharge threshold sets here to 0.5,

Which is exactly equivalent to:

$$\text{spiketrain}\,(t) = \begin{Bmatrix} 1 \text{ if } \mathrm{U}(t) \times \mathrm{Signal}_{\text{signal}>0}(t) \times P_{\text{refract}}(t) > \mathrm{Thres} \\ 0 \text{ if } \mathrm{U}(t) \times \mathrm{Signal}_{\text{signal}>0}(t) \times P_{\text{refract}}(t) > \mathrm{Thres} \end{Bmatrix} \quad (2)$$

where

$Signal_{signal>0}(t)$ is the positive part of the input signal at the output of an auditory filter.

As an intern noise *U* has been added, each signal is passed 300 times in the simulation model to estimate the distribution of the periodicities of the spiketrain. Moreover, for each signal, a single auditory filter output, located in the passband of the input signal and centred on the carrier frequency (f_c), has been used.

2.2 Stimuli

The stimuli used by Santurette and Dau (2011) and by Oxenham et al. (2011) have been generated and processed through the simulation. In Santurette and Dau (2011), the signals were generated by multiplying a pure tone carrier with frequency f_c with a half-wave rectified sine wave modulator with modulation frequency f_{env} and low-pass filtered by a 4th order Butterworth filter with cut-off frequency of $0.2 \times f_c$. All signals were generated at 50 dB SPL and mixed with a TEN noise at 34 dB SPL per ERB. All f_c and f_{env} values are indicated in Fig. 2. When f_c and f_{env} are not multiple from each other, this manipulation produce a shifted complex. In Oxenham et al. (2011), harmonic complex tones at various F0 values (indicated in Fig. 3) were generated by adding in random phase up to 12 consecutive harmonics, beginning on the sixth. Harmonics above 20 kHz were not generated. All harmonics were generated at 55 dB SPL per component and all signals were embedded in a broadband TEN noise at 45 dB per ERB. A shifted version of each harmonic complex tone was also generated by shifting all components of the complex tone by an amount of $0.5 \times$ F0.

3 Results

The outputs of the simulation provide the distributions of the temporal periodicities of the spiketrain. This is supposed to predict the perceived pitch evoked by the signal.

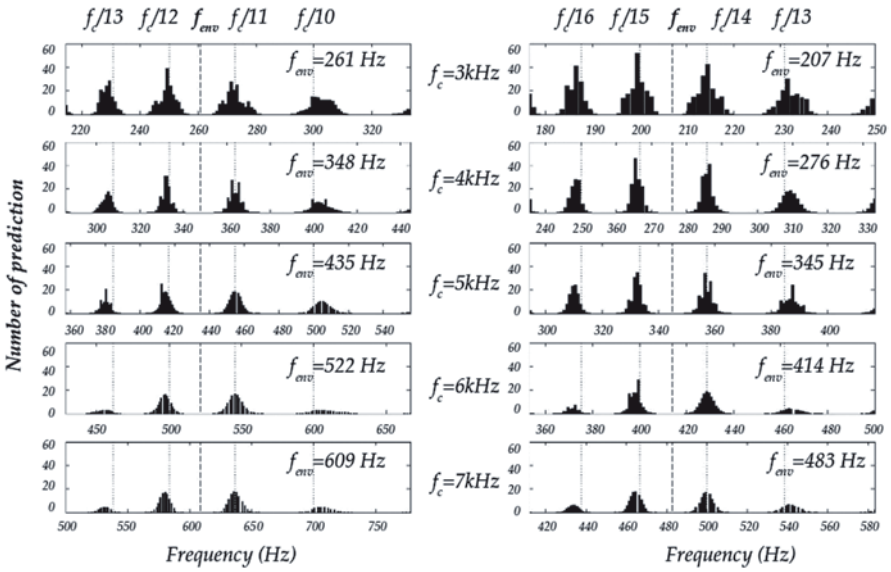

Fig. 2 Outputs of the simulation fed with the signals used in Santurette and Dau (2011). The distributions of pitch estimation are always related to TE × TSF_{period} and never on TE_{period}. This is closely consistent with the data reported in Figs. 4 and 6 in Santurette and Dau. (2011)

The results of the simulation plotted in Fig. 2 evidence that, as in Santurette and Dau (2011), the predicted pitch values are always related to TE × TSF_{period} and never related to TE_{period}. This is strongly consistent with the results reported by the authors.

The results of the simulation plotted in Fig. 3 are strongly consistent with Oxenham et al. (2011). Using a plausible refractory period of 2500 Hz, the simulation is able to extract the TE × TSF_{period} even if the TSF_{period} are too fast to be correctly encoded. The decrease in performances reported in Oxenham et al. (2011) when

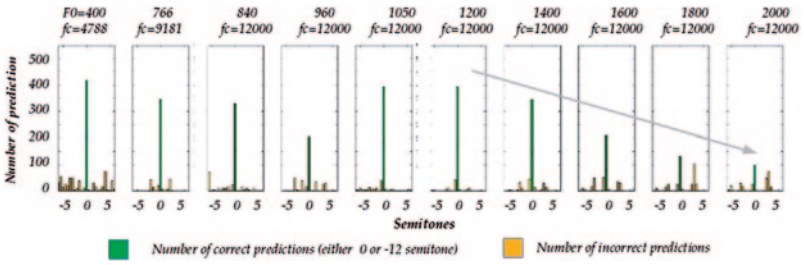

Fig. 3 Outputs of the simulation at the auditory filters centred on fc, fed with the signals used in Oxenham et al. (2011). The distributions of pitch estimation are closely consistent with the data reported in Fig. 2B in Oxenham et al. (2011). The *gray arrow* indicates the decrease in the number of predictions when increasing the resolvability of the complex by increasing the F0

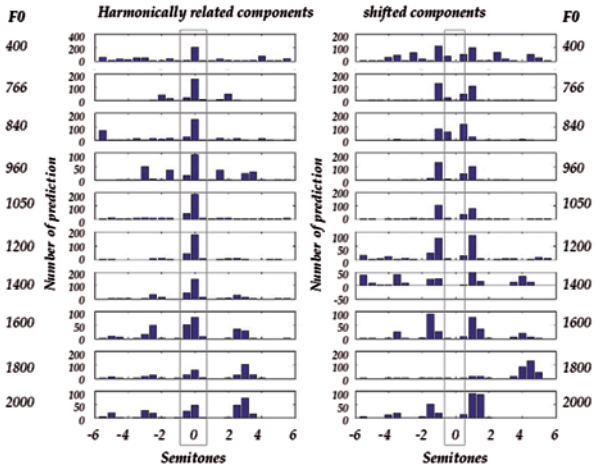

Fig. 4 Output of the simulation fed with the signals used in Oxenham et al. (2011) with harmonic complex tones (*left* column) and shifted complex tones (*right* column). The distributions of pitch estimation are always related to TE × TSF_{period} which predict a peak centred in the *gray rectangle* on the *left* and peaks on either side of the *gray rectangle* on the *right*. This is closely consistent with the data reported in experiment 1 and 2 in Oxenham et al. (2011)

increasing the F0 from 1200 Hz to up to 2000 Hz is also simulated. This decrease is probably explained by a decrease in resolvability (less and less interactions between harmonic components) when increasing the F0.

Finally, the results of the simulation plotted on the right column of Fig. 4 are strongly consistent with the results found with the signals used by Santurette and Dau (2011), and also evidence that the predicted pitch values are always related to TE × TSF_{period} and never related to TE_{period}.

As a control, the effect of threshold value used in the simulation has been tested with one complex tone having a F0 equal to 1400 Hz (Fig. 5). Varying the threshold value have some important incidence on the predictions. Using a threshold below 0.3 does not provide reliable periodicities estimations. Using a threshold from 0.4 to 0.8 provides reliable and consistent estimations. Using a high threshold value (over 0.9) prevents to report any periodicities. Using 0.5 as in the previous simulations appears then to be a good compromise. It is worth noting that such a threshold model is physiologically plausible and could be related to the thresholds of the auditory nerve fibres previously described in the literature (Sachs and Abbas 1974).

4 Discussion

4.1 Conclusions

These stimulations have a double interest.

First, this evidence that any perceptual effect that is empirically evidenced between harmonic complex tones and shifted complex tones should not been interpreted as a pure effect of fine structure. In fact, the pitch evoked by a shifted complex is based on interaction cues between envelope and fine structure (TE × TSF_{period}).

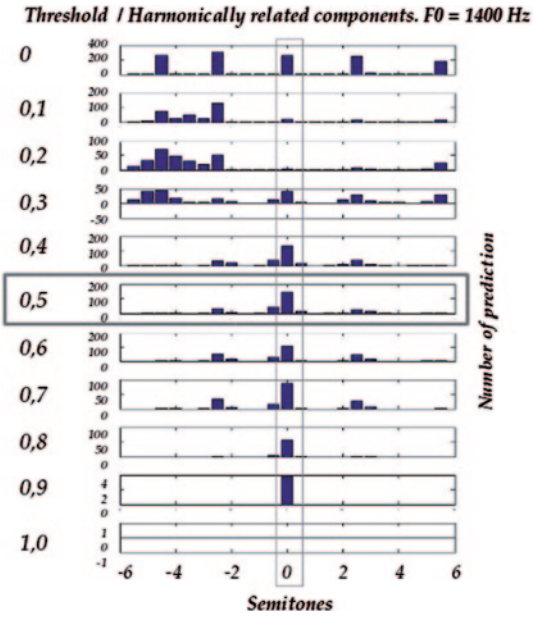

Fig. 5 Output of the simulation fed with a single harmonic complex tone (F0 = 1400 Hz) used in Oxenham et al. (2011). Effect of threshold value from 0 to 1 on periodicity estimations

Using these signals to tease apart temporal envelope cues from temporal fine structure cues is then a conceptual error. This impaired the conclusion that the pitch of unresolved complex tones is based *only* on fine structure information.

Second, when thinking about pitch perception of unresolved complex tones in terms of interaction between envelope and fine structure, it appears that the limitation of phase locking is probably much less critical than when thinking in terms of fine structure only. In fact, it seems clear that there is no need to encode every phases of the signal to encode the most intense phases located nearby an envelope maximum. This explains that the simulation can extract a periodicity related to pitch when the carrier frequency is over 10 kHz (Fig. 3).

4.2 Limitations

First, the current simulations are not a physiologically-based model of pitch perception and the refractory period which is used here does not accurately describe the physiological constraints for the phase locking. Some further works that would use realistic models of the auditory periphery should be used for further explorations.

Second, this simulation does not explain all the data reported in the literature about pitch perception. For example, experiment 2c in Oxenham et al. (2011) reports some pitch perception using dichotic stimulations with even-numbered har-

monics presented on the right ear and odd-numbered harmonics presented to the left ear. This experimental manipulation increases the resolvability of the signals and prevents our simulation to extract any temporal periodicities and then to predict some pitch perception.

Acknowledgments This work was supported by institutional grants ("Centre Lyonnais d'Acoustique," ANR-10-LABX-60) and (Contint—Aïda, ANR-13-CORD-0001). I thank Andy Oxenham and Etienne Gaudrain for constructive and interesting comments about this work.

References

De Boer E (1956) Pitch of inharmonic signals. Nature 178(4532):535–536

Irino T, Patterson RD (2006) A dynamic compressive gammachirp auditory filterbank, IEEE Trans Audio, Speech, Lang process 14(6):2222–2232

Oxenham AJ, Micheyl C, Keebler MV, Loper A, Santurette S (2011). Pitch perception beyond the traditional existence region of pitch. Proc Nat Acad Sci U S A 108(18):7629–7634. http://doi.org/10.1073/pnas.1015291108

Sachs MB, Abbas PJ (1974) Rate versus level functions for auditory-nerve fibers in cats: tone-burst stimuli. J Acoust Soc Am 56(6):1835–1847

Santurette S, Dau T (2011). The role of temporal fine structure information for the low pitch of high-frequency complex tones. J Acoust Soc Am 129(1):282–292. http://doi.org/10.1121/1.3518718

Smith ZM, Delgutte B, Oxenham AJ (2002). Chimaeric sounds reveal dichotomies in auditory perception. Nature 416(6876):87–90. http://doi.org/10.1038/416087a