



Detailed Multidimensional Analysis of Our Acoustical Environment

Marián Képesi

Signal Processing and Speech Communication Laboratory, Technical University of Graz, A-8010 Graz, Inffeldgasse 12, Austria,
e-mail: kepesi@tugraz.at,

Luis Weruaga

Austrian Academy of Sciences, A-1220 Vienna, Donau-City Str. 1., Austria, e-mail: weruaga@ieee.org

Edward Schofield

Forschungszentrum Telekommunikation Wien, A-1220 Vienna, Donau-City Str. 1., Austria, e-mail: schofield@ftw.at

The most often-used time-frequency analysis tool, the Short-Time Fourier Transform, suffers from blurry harmonic representation when analysing acoustic sources with changing frequency. This phenomenon is always present in the analysis of underwater and geophysical signals, the sounds of bats, whales, birds and other animals, engine noises, and human conversations. For time-frequency analysis of real-world acoustic signals like these this paper introduces a technique based on the Short-Time Harmonic Chirp Transform (STHChT) with a noise-level-independent feedback-like estimation of frequency changes. The basis of this adaptive transform comprises quadratic chirps that follow the frequency trajectory segment by segment. The frequency tracking method is based on the harmonic structure of the source, derived from its spectral representation, whose performance is in turn strongly enhanced by the use of the STHChT. This combination of analysis tools offers a more precise time-frequency representation of acoustic signals than state-of-the-art time-frequency analysis techniques. Comparative evaluation results between the proposed STHChT and popular time-frequency techniques reveal an improvement in time-frequency localisation and a finer spectral representation.

1 Introduction

The Short-Time Fourier Transform is the most popular time-frequency analysis tool applied in acoustic signal processing. This transform yields decent spectral representations of periodic signals, but suffers from blurry harmonic representations of quasi-periodic signals that undergo changes in frequency. In the real world such frequency variations can be relatively high in comparison with the analysis window length, leading to smearing across bins in the frequency or time domain. For this reason time-frequency signal analysis is still a topic of active research [1, 2, 3].

This paper describes a method to achieve accurate time-frequency representations of acoustic signals despite variations in their frequency. Since chirpy signals are fundamental in nature, precise multidimensional representations of such non-stationary signals are of great importance. Bat and whale signals have been the subject of much research, often using chirplets to analyze their time-frequency structure. Haykin *et al.* [4] proposed the so-called Chirplet transform; Mihovilovic *et al.* [5] considered Chirplet banks; Baraniuk *et al.* [6] considered warped wavelets. The choice of the appropriate basis according to the time-frequency characteristics of the analyzed signal is still an open research topic; recent research is based on search methods [7, 8], approximating the chirpy signal with a linear Gaussian chirp and searching for the most plausible chirp rate (relative frequency

change) among a range of candidates.

We have previously shown in [9, 10] that the Short-Time Harmonic Chirp Transform can offer a very detailed time-frequency representation of speech signals, especially those with changing pitch, as often occurs in normal intonation. This paper describes tests applying this transform to a variety of sounds from nature and our everyday acoustic environment.

In this paper the basis of the Short-Time Harmonic Chirp Transform will comprise quadratic chirps whose chirp rate is reestimated for each analysis segment according to the pitch evolution. The use of a chirp basis corresponds to the study published in [11].

We estimate the pitch trajectory segment by segment by the spectral gathering method described below and in [12]. This method provides estimates of common harmonicities in time-frequency space, which means that besides the exact pitch evolution, further harmonicities present in the signal are derived from the short-time spectrum. Precise frequency estimation is the key to achieving the potential of the adaptive Short-Time Harmonic Chirp Transform.

The paper is organised as follows: Section 2 introduces the spectral gathering method (SGM) for precise frequency estimation; Section 3 describes the STHChT as a tool for time-frequency analysis; and Section 4 discusses evaluations on real recordings from our environment.

2 Harmonicity analysis by spectral gathering

During Lombard speech, changes in the formant positions and harmonic structure allow one to become more understandable in the presence of background noise. Likewise, the harmonic structure of acoustic sources is an important clue for the human perception of different sources in a noisy environment and for distinguishing between a mixture of these sources. The main idea of our time-frequency analysis tool is to aggregate energies over all the harmonicities belonging to the same pitch value. More precisely, for each hypothesised pitch value f_0 in some range $[f_{\min}, f_{\max}]$, the algorithm sums over the energy of its higher harmonic components. We collect the log-amplitudes of harmonics, since the human auditory system is sensitive to energy on a logarithmic scale, into a ‘gathered log spectrum’:

$$\rho_0(f_0) = \frac{1}{H} \sum_{h=1}^H \log_{10} S(hf_0), \quad (1)$$

where $S(f)$ is a spectral representation of the speech segment under analysis and H is the number of harmonics considered. After evaluating this energy reassignment, we estimate the candidate pitch value F_0 for the analysed segment as the position of the highest peak:

$$F_0 = \arg \max_{f_0} \rho(f_0). \quad (2)$$

We compute the gathered log spectrum (1) for only a restricted range of frequencies $[f_{\min}, f_{\max}] = [50, 500]$ for speech, $[200, 1000]$ for cats, $[500, 3000]$ for birds, etc.

Applying this kind of estimation method faces two problems in real applications: first, our discrete short-time spectral representation is of finite resolution (often above 30Hz per bin), and second, there are additional peaks present in the new gathered spectrum (1) corresponding to multiples and halves of the real pitch value.

An effective workaround for the first problem is to use linear approximation of spectral bins, estimating the energy of a conjectured harmonic component at some frequency f_0 from even a low-resolution spectral representation by linear approximation:

$$S(f_0) = (1 - d) S(\hat{k}_0) + d S(\hat{k}_0 + 1), \quad (3)$$

where $S(k)$ is the short-time spectrum of the analysed segment,

$$\begin{aligned} k_0 &= \frac{f_0}{f_S} \cdot N, \\ \hat{k}_0 &= \text{floor}(k_0), \end{aligned} \quad (4)$$

and

$$d = k_0 - \hat{k}_0. \quad (5)$$

By visual inspection of the gathered log-spectrum (1) we can define the notion of fundamental frequency: we decide that frequency f is the fundamental frequency F_0 if none of its simple fractions f/l for integers $l > 1$ is of a comparable gathered level, that is, if $\rho_0(f) \approx \rho_0 \gg (f/l)$ for all l . We penalise hypothetical harmonic peaks not fulfilling this condition with the following modification to the gathered log spectrum:

$$\rho(f) = \rho_0(f) - \max_l \{ \rho_0(f/l) \}. \quad (6)$$

All of our experiments described here use this second form of energy reassignment to remove anomalies. For a more detailed explanation of this method of pitch extraction see [12].

3 The Short-Time Harmonic Chirp Transform

Figure 1 shows log spectra of voiced female speech yielded by the Short-Time Fourier Transform (STFT) and the Short-Time Harmonic Chirp Transform (STHChT) [9]. Notice the smearing across frequency in the Fourier spectrum (Figure 1a) resulting from the changing pitch, and how the chirp spectrum has cleaner peaks corresponding to the harmonics.

The STHChT can be understood as a generalised STFT without the assumption of constant pitch within each analysis frame. If the Fourier transform of a speech segment is defined as

$$S(k) = \sum_{n=0}^{N-1} x[n] w[n] \exp(-j 2\pi k n/N) \quad (7)$$

(where $x[n]$ is the discrete signal under analysis, $w[n]$ is a sliding analysis window of length N samples, n is the discrete time index, and k discrete frequency), by adding an extra parameter α representing the normalised frequency variation rate and by replacing the sinusoidal basis with one of linearly varying complex exponentials (or chirps) defined as

$$\xi(m, k, \alpha) = \exp(j 2\pi k m (1 + \alpha (m - N))/N) \quad (8)$$

we reach the definition of the Harmonic Chirp transform:

$$C_x(n, k, \alpha_n) = \sum_{m=0}^{N-1} x[m + nM] w[m] \xi(m, k, \alpha_n)^* \quad (9)$$

Here the superscript $*$ denotes complex conjugation and α_n is the discretised time-variant chirp-rate for the n th segment.

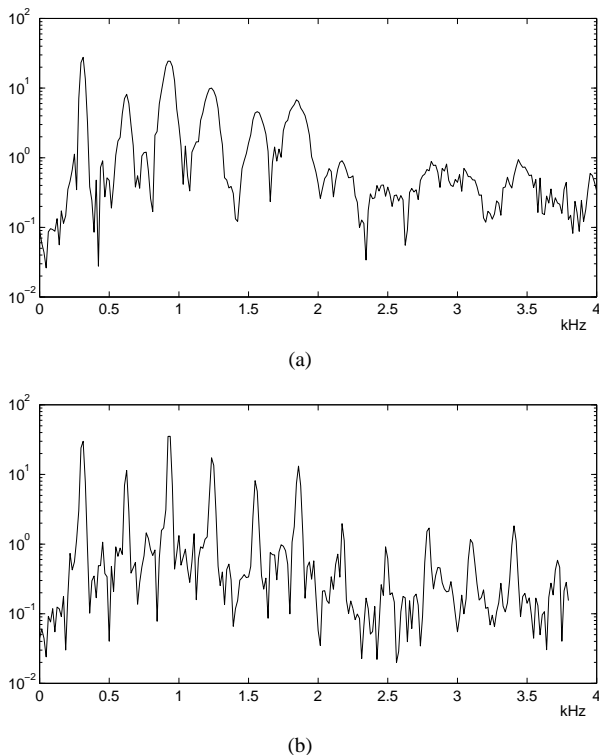


Figure 1: *Fourier and chirp log-spectra of a voiced female speech segment with slightly changing pitch.*

We assume that the instantaneous frequency of the harmonic component to represent in detail changes linearly during the analysis interval, rather than remaining constant. This yields a more precise time-frequency representation, but assumes that the pitch-change rate of the segment is known ($\alpha_n \approx \gamma_n$).

Precise time-frequency resolution with the STHChT in on-line applications requires accurate estimation of the chirp rate α_n . This parameter is derived from the average frequency variation of the main signal components in the n th segment, normalised by the mean frequency value and the segment length:

$$\alpha_n = \frac{\Delta F_0}{F_0 M} = \frac{2(F_n - F_{n-1})}{M(F_n + F_{n-1})} \quad (10)$$

Here n is the discrete time-index, and F_n and F_{n-1} are the mean frequency F_0 calculated from the actual and previous segments. The chirp-rate calculated from (10) holds at $m = 0$ (the start of the analysis segment) and could differ slightly from the pitch-change rate at $m = N - 1$.

As we have mentioned, the STHChT is a general formulation of the STFT with one additional parameter, the chirp rate. If this rate is set to zero the STHChT reduces to the standard STFT. Additionally, when applying this transform with $\alpha \neq 0$, the active spectral area is restricted due to aliasing arising from the sweeping

basis waves. A more precise representation is achievable than with the STFT since the basis functions can exactly match the pitch-change rate of the speech signal. Figure 1b demonstrated this on a speech signal, but the next chapter will show that the same holds true for other environmental sounds.

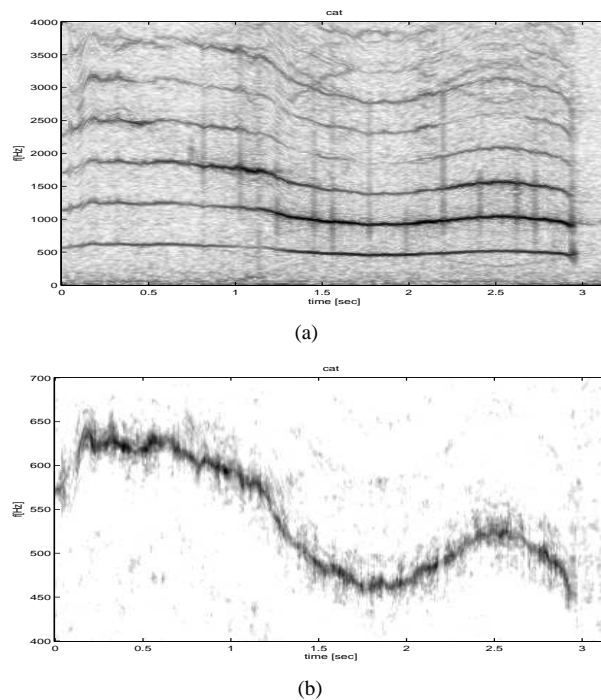


Figure 2: *Harmonicity analysis of a cat's voice: (a) Short-Time Harmonic Chirp representation, (b) time evolution of harmonicities*

4 Results and Discussion

This section presents a comparative analysis of time-frequency representations produced by the STHChT and the STFT. The sound recordings are available from <http://edschofield.com/resources/fa2005>.

The acoustic signals were sampled at 8 and 11 kHz. A Hamming window was applied of length 256 samples with a constant stepsize $M = 64$. The weighted frames were zero-padded before applying the transforms.

Note that, although longer windows normally yields more smeared spectral bins with the STFT, longer windows can be applied with the STHChT without a problem, yielding much more detailed time-frequency representations.

The effectiveness of the STHChT over the classical spectrogram-type representation is clearly visible in case of different mammals sounds (see Figures 3 and 4.) Figure 3 shows spectrograms of a cat's voice using the STFT and STHChT and their corresponding gathered spectra. Notice the improved clarity offered by the STHChT.

The benefit of the spectral gathering method in separating two sources with overlapping time-frequency structure is demonstrated in Figure 6. However, for sounds with only one or few harmonics, like the left side of Figure 4, the gathered spectrum differs little from the spectral representation, and is not shown. This implies that a nonlinear distortion might be performed in the cochlea to enrich the spectral representation of sounds for pitch tracking.

5 Summary

This paper has described an application of the Short-Time Harmonic Chirp Transform, together with the spectral gathering method (SGM) for pitch extraction, to the time-frequency analysis of several natural and environmental sounds. Although these methods were originally designed with a focus on speech analysis and pitch estimation in noisy environments, this paper has shown that they have potential for the analysis of acoustic sources like underwater sounds, bird songs, and mammal sounds. The frequency resolution of the method proposed is higher than the that provided by the Fourier transform for frequency-varying signals, since the chirpy basis allows the use of longer analysis windows. The main advantage of this method over other analysis methods is with quasi-periodic signals with variations in the mean frequency rate. This paper has discussed only segment-wise pitch tracking; ensuring continuity of pitch estimates across segments remains an open topic for future research.

6 Acknowledgments

This work was partially supported by the Mistral project at SPSC, TU Graz and the K-plus program of the Austrian government.

References

- [1] L. Cohen, *Time-frequency analysis*. Englewood Cliffs, NJ: Prentice Hall, 1995.
- [2] P. Flandrin, *Time-Frequency/Time-Scale Analysis*, Academic Press, San Diego, 1999.
- [3] T. F. Quatieri, *Discrete-time speech signal processing*. Upper Saddle River, NJ: Prentice Hall, 2002.
- [4] S. Mann, S. Haykin, "The chirplet transform: physical considerations", *IEEE Trans. Signal Proc.*, pp 2745–2761, Nov., 1995.
- [5] D. Mihovilovic, R.N. Bracewell, "Whistler analysis in the time-frequency plane using chirplets," *J. Geophys. Res.*, vol. 97, no. A11, pp. 17199-17204, Nov. 1992.
- [6] R.G. Baraniuk and D.L. Jones, "Warped wavelet bases: unitary equivalence and signal processing," *Proc. IEEE ICASSP*, pp. 320-323, Minneapolis, April 1993.
- [7] Q. Yin, S. Qian and A. Feng, "A fast refinement for adaptive gaussian chirplet decomposition," *IEEE Trans. Signal Processing*, vol. 50, no. 6, pp. 1298–1306, June 2002.
- [8] J.C. O'Neill, P. Flandrin, "Chirp hunting," *Proc. IEEE Int. Symp. on Time-Freq and Time-Scale Anal.*, pp. 425-428, 1998.
- [9] L. Weruaga, M. Képesi, "Speech Analysis with the Short-Time Chirp Transform," *Proc. of Eurospeech 2003*, Geneva, CH, Sept. 2003, pp. 53–56,
- [10] L. Weruaga, Képesi, M., "Speech analysis with the fast chirp transform", *Proc. EUSIPCO*, Vienna, AT, 2004, pp. 1011–1014.
- [11] E. Mercado, C. E. Myers, and M. A. Gluck, "Modeling auditory cortical processing as an adaptive chirplet transform," *Neurocomputing*, 32-33, pp. 913-919, 2000.
- [12] M. Képesi, L. Weruaga, "High-Resolution Noise-Robust Spectral-based Pitch Estimation", submitted to *Eurospeech 2005*, Lisboa, Portugal.

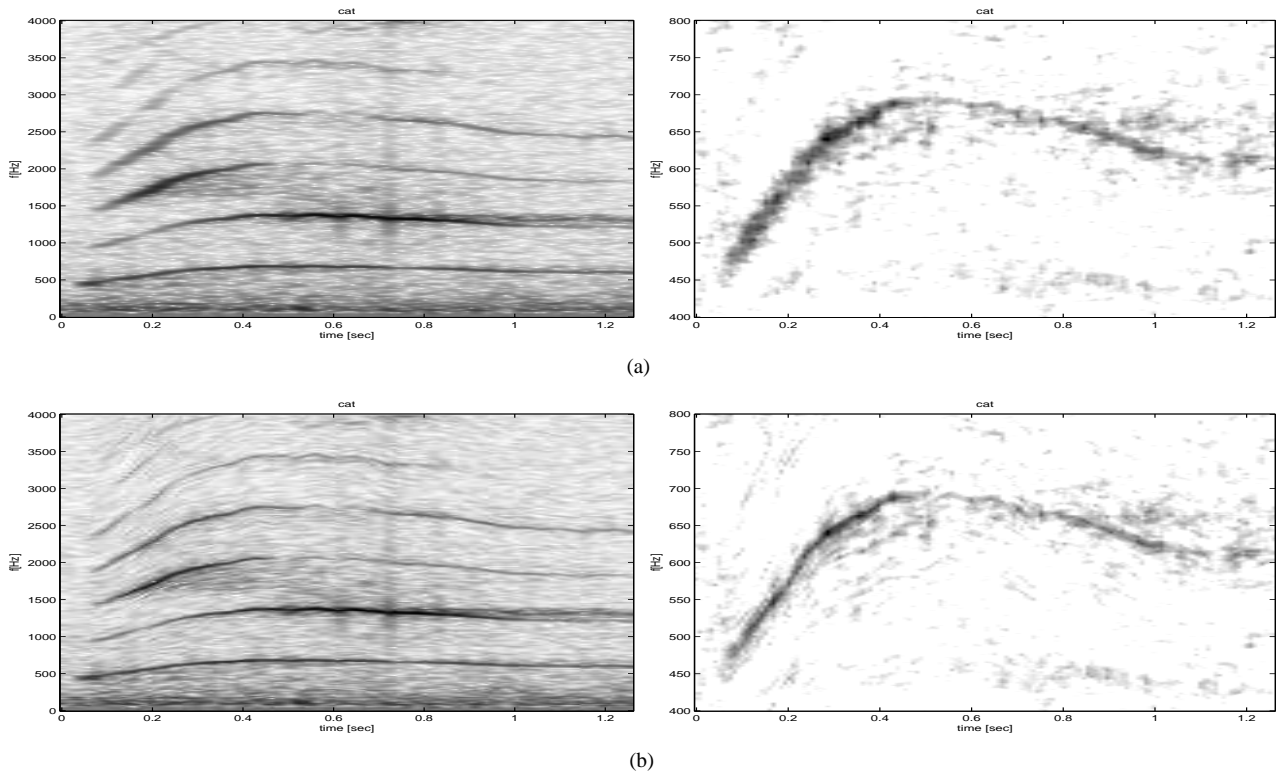


Figure 3: *Harmonicity analysis of a cat's voice: (a) Short-Time Fourier representation, and STFT-based time evolution of harmonics (b) Short-Time Harmonic Chirp representation, and STHChT-based time evolution of harmonics*

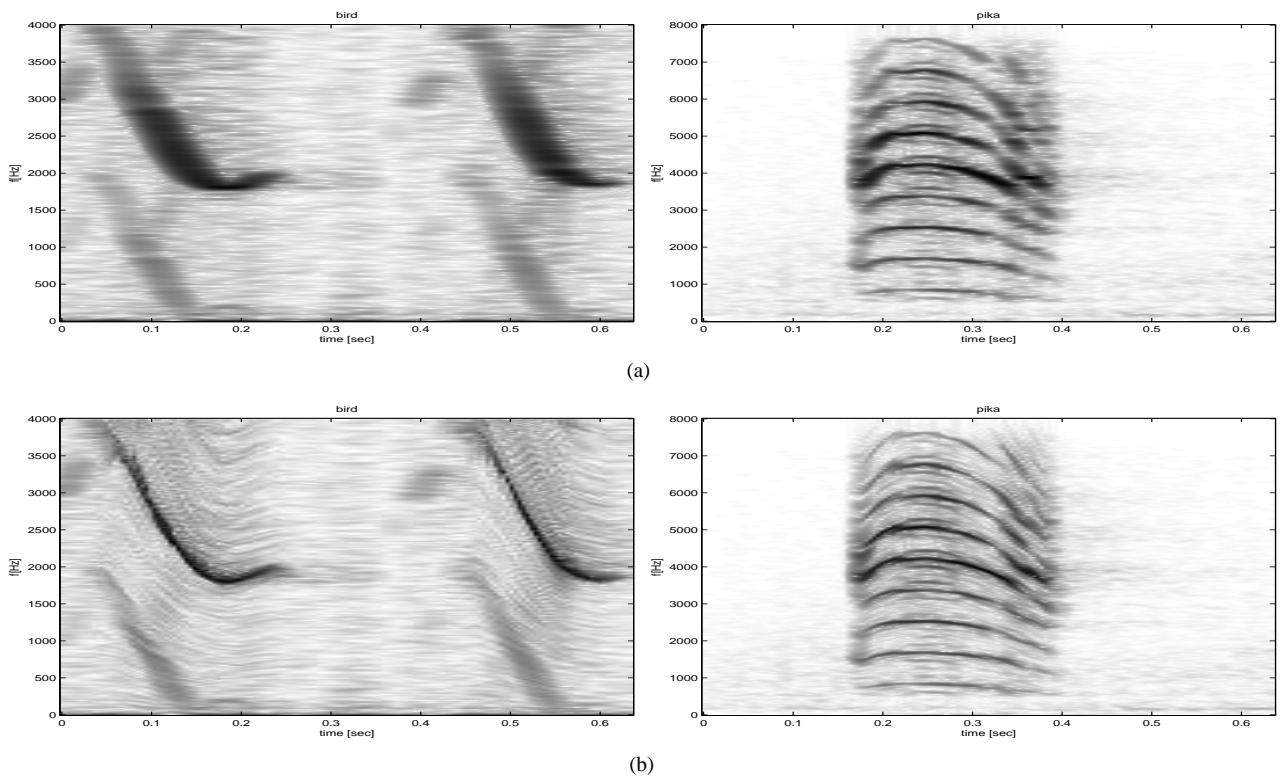


Figure 4: *STFT vs. STHChT analysis: (a) Short-Time Fourier representation of a bird song, (b) its Short-Time Harmonic Chirp representation.*

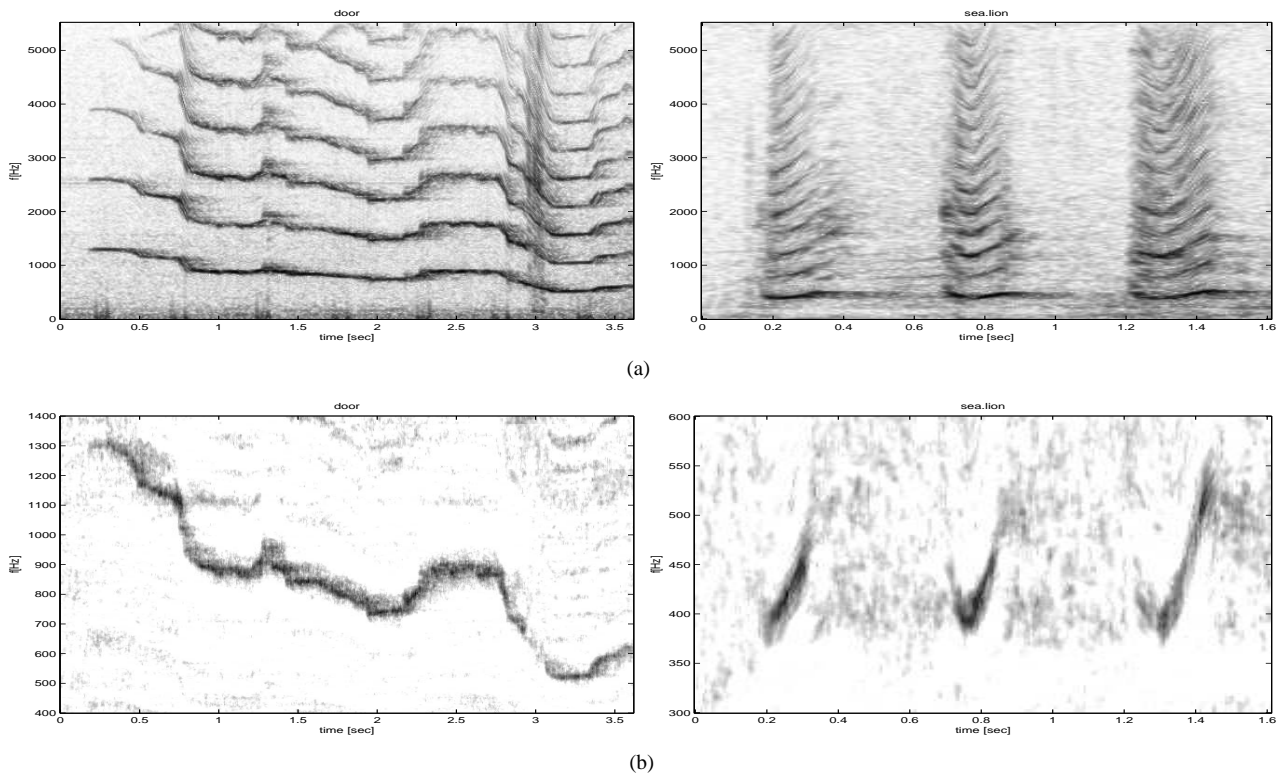


Figure 5: Short-Time Harmonic Chirp representation and time evolution of harmonicities: (a) closing door, (b) sea-lion.

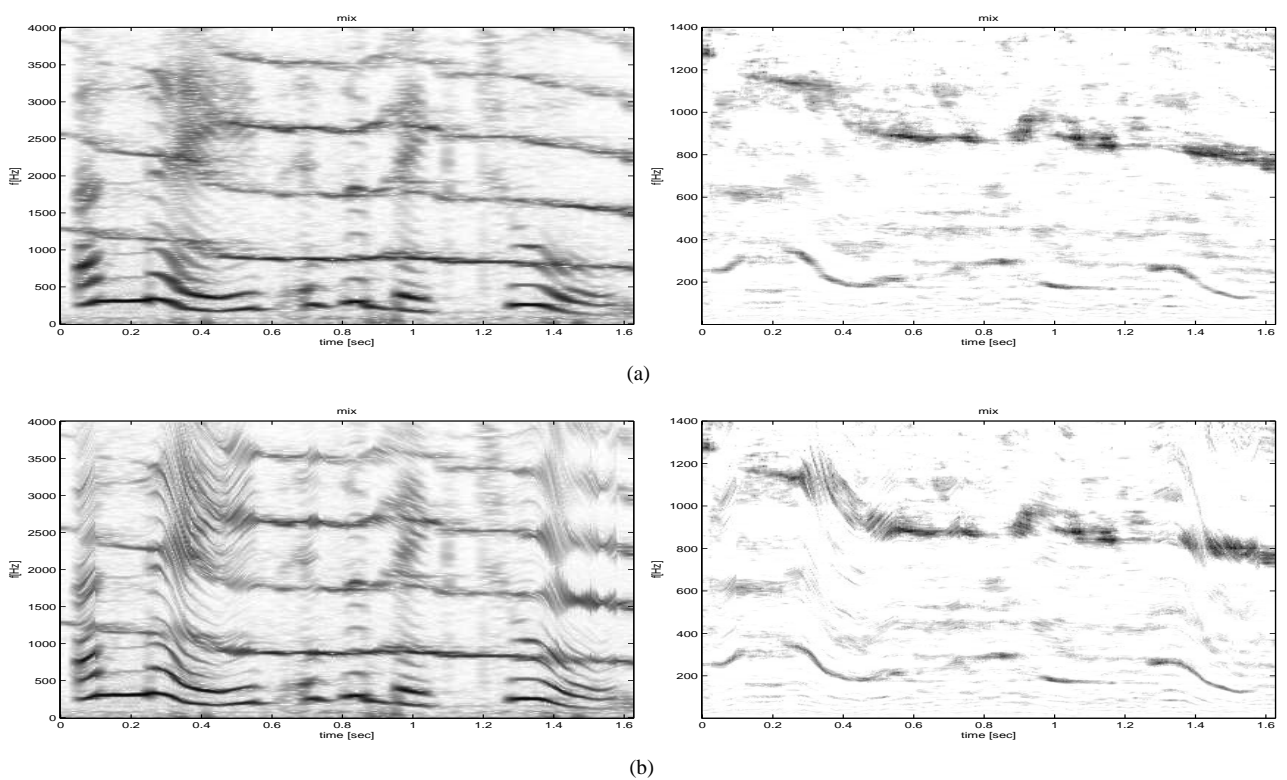


Figure 6: Mixture of a closing door and female speech: (a) Fourier analysis and spectral gathering; (b) chirp analysis and spectral gathering