Time-scaling Phase Vocoders

Audio Signal Processing

Yoach Lacombe
MVA 21/22
ENS Paris-Saclay

March 29, 2022

Sommaire

Introduction

Time-scale modification (TSM)

Time-scale algorithm aims at stretching the length of an audio signal while preserving its pitch and timbre. There are two main paradigms for TSM [Driedger, 2016].

- **Time-domain-based vocoders** modify the audio via the time domain. Mostly based on Overlapp-Add(OLA).
- **Frequency-domain-based vocoders** modify the audio via the phase of the short-time Fourier Transform (STFT). Mostly based on Phase-Vocoder (PV).

Some notation

- N denotes the length of the windowed signal of the STFT and F_s is the framerate of the input audio signal x
- R_a and R_s respectively denotes the analysis and the synthesis hop sizes.
- h is a window of size N, here the hanning window.

The basic pipeline

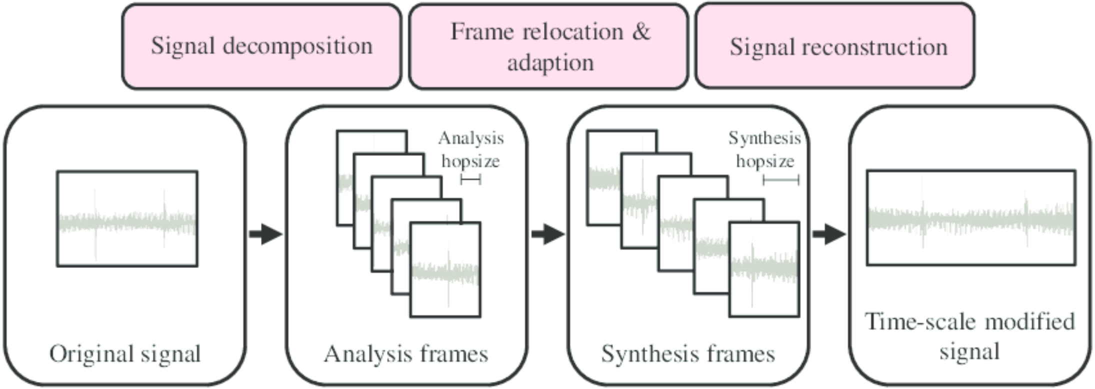

Figure: TSM process. Analysis, Modification, Synthesis. From [Driedger, 2016].

Analysis

Analysis stage

Basically, analysis applies STFT to the input signal x.

$$X(m,k) = \sum_{n=0}^{N-1} h(n)x(n+mR_a)e^{-2i\pi kn/N} \tag{1}$$

On the bins

(m,k) is a time-frequency bin associated to the time mR_a/F_s and frequency kF_s/N.

Synthesis

The **synthesis** stage typically uses the inverse (in the least square sense) STFT. Short-time signals $y_m(n)$ are obtained by computing the inverse FFT of Y. These signals are then weighted by a synthesis window h (typically the hanning window) and overlapp-add by a synthesis hop R_s to compute the output signal y.

$$y_m(n) = \frac{1}{N} \sum_{k=0}^{N-1} Y(m,k) e^{2i\pi kn/N}$$

$$y(n) = \sum_{m=-\infty}^{\infty} h(n - mR_s) y_m(n - mR_s)$$

PV-TSM

Recurrence

With Φ_* denoting the phase, at time m,

$$\Phi_Y[m,:] = \Phi_Y[m-1,:] + R_s IF(m) \tag{2}$$

where $IF(m) = \Omega + \frac{1}{R_a}[\Phi_X[m,:] - \Phi_X[m-1,:] - R_a\Omega]_{2\pi}$ with $\Omega = \{k\frac{2\pi}{N}\}_{k\in\mathbb{N}_N}$

Initialisation

$$|Y| = |X| \tag{3}$$

$$\Phi_Y[0,:] = \beta\Phi_X[0,:] \tag{4}$$

Here, β is a parameter that is usually set to 1 which [Laroche, 1999] brings to solve the phasiness issue.

Issues and proposed improvements

Issues with PV-TSM

Main issues are:

- Transient smearing - loss of percussiveness
- Phasiness - the speaker seems to be away from the mic

Phasiness is identified to be caused by lack of vertical phase coherence.

Some proposed improvements

To solve this vertical phase coherence issue, [Laroche, 1999] proposes phase-locking.
Other improvements could be:

- Resetting the output frames to the input frames every D processed frames.
- Dynamically change the phase either horizontally or vertically according to gradient.
- Automatically adapt the time-frequency resolution to analyse and resynthesize.

Phase-locking

Vertical phase coherence: a sinusoidal component may affect multiple adjacent frequency bins of a single analysis frame.

Assumption

A frame's magnitude is representative of a particular sinusoidal component and that the surrounding bins with lower magnitude are affected by this very same sinusoidal component

A naive first approach - *loose phase-locking*

By simply computing a vertical rolling sum on the resulting STFT, the bins of higher amplitude dominate their neighbors.

$$Y_{new}[:, k] = Y[:, k-1] + Y[:, k] + Y[:, k+1]$$

Identity Phase-locking

Identity phase-locking [Laroche, 1999] consists in the following steps, for each analysis frame.

1. Identify the peaks which are identified as bins where the magnitude is larger than the 4 nearest neighbors.
2. Compute the synthesis frame phase of each peak according to Equation [2].
3. Identify the closest channels to each peak.
4. For a channel k and its closest channel k_l, update the synthesis phase such that:

$$\Phi_Y[m,k] = \Phi_Y[m,k_l] + \Phi_X[m,k] - \Phi_X[m,k_l]$$

Issues

While highly improving the quality of the TSM, it still suffers from transient smearing and to recurring interruptions (when harmonic signal happens at the same time as transient smearing).

HPS-based TSM [Driedger, 2014]

With regards to the remaining issues of phase-locked PV-TSM, we remark that:

- PV-TSM is particularly adapted for the harmonic part of a sound
- OLA (overlapp-add approach) is particularly adapated for the percussive part of a sound.

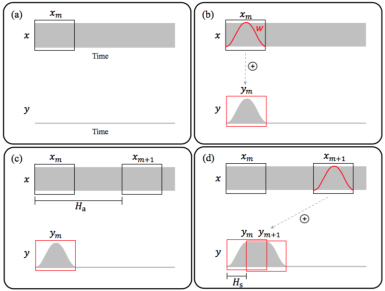

Figure 3. The principle of TSM based on overlap-add (OLA). **(a)** Input audio signal x with analysis frame x_m. The output signal y is constructed iteratively; **(b)** Application of Hann window function w to the analysis frame x_m resulting in the synthesis frame y_m; **(c)** The next analysis frame x_{m+1} having a specified distance of H_a samples from x_m; **(d)** Overlap-add using the specified synthesis hopsize H_s.

Figure: OLA principle explained (from [Driedger, 2016]).

HP-based TSM - 2

1. Separate harmonic and percussive components by applying vertically and horizontally median-filters.
2. Apply OLA to the percussive component and PV-TSA to the harmonic component.
3. Add back the two resulting signals.

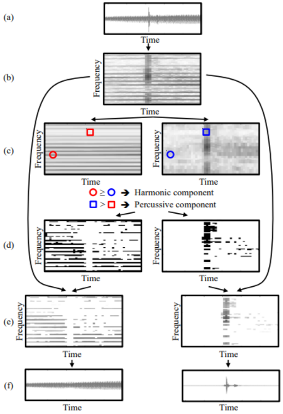

Figure: HPS approach's pipeline (from [Driedger, 2016]).

Informal observation - 1

Overall observation

- Despite the transient smearing and some unpleasant interruptions, the Identity Phase-Locked PV stays a really strong algorithm with satisfactory results. It efficiently solves the phasiness issue.
- HPS-TSM performs really well too. As expected it solves the transient smearing and most of the unpleasant interruptions are cleared.

On hyperparameters

- **Lengths of the median filters**: Important parameters. The most robust values I found are 100 for time spectrogram and 25 for frequency spectrogram.
- **N:** As expected, crucial as it is a tradeoff between time and frequency resolutions. Set to correspond to 100 ms. $N = 0.1F_s$.

Informal observation - 2

On hyperparameters - **Hop Sizes**

- Most of the papers I have read recommend setting $R_s = N/2$ and thus $R_a = R_s/\alpha$.
- However, R_s and R_a become too large. The analysis STFT loses information.
- Might explain some of the inconsistencies of [Laroche, 1999] such as the importance of β and the Scaled Phase-Locked results.
- With hop sizes set to reasonable values (ex. $R_a = 128$ and $R_s = \alpha R_a$), every algorithms sound better with every sound samples.

On HPS

Even without applying TSM, the separation of harmonic and percussive components work particularly well. See the following slide for example.

HPS example

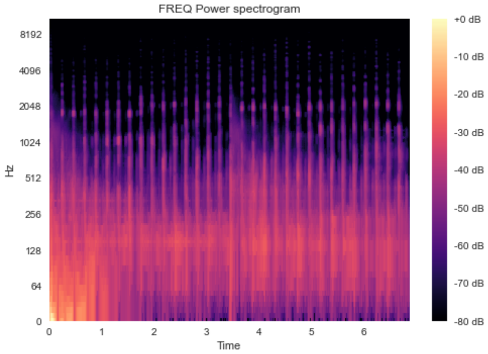

Figure: Frequency component after applying the percussive mask.

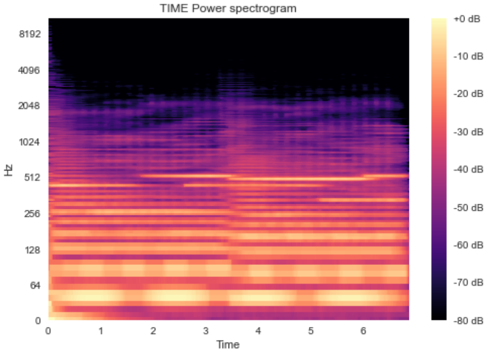

Figure: Time component after applying the harmonic mask).

Conclusion

- Some hyperparameters are crucial to hear pleasant results. Moreover, the values recommended in the literature appeared are not suited for the time-stretching task.
- **Limitation:** Should have tested with more complex polyphonic sounds. I haven't test the algorithms with polychannel audio signals as well.
- To truly evaluate methods, we should conduct formal listening test. The consistency measure proposed in [Laroche, 1999] which compares Y to the STFT of y appears not to be used in other papers and is not based on any valid background.

References

Jonathan Driedger and Meinard Muller.(2016)
A review of time- scale modification of music signals.
Applied Sciences, 6(2)

Laroche and M. Dolson (1999)
Improved phase vocoder time- scale modification of audio
7(3):323–332, 1999

Jonathan Driedger, Meinard Muller and Sebastian Ewert. (2014)
Improving time-scale modification of music signals using harmonic-percussive separation.
21:105–109, 2014

The End