ON COMPARISON OF PHASE ALIGNMENTS OF HARMONIC COMPONENTS

Wen Xue[1], *Lou Xiaoyan*[1], *Mark Sandler*[2]

Samsung Electronics[1]
Beijing, China

Queen Mary University of London[2]
London, UK

{xue.wen,xiaoyan.lou}@samsung.com[1],
mark.sandler@eecs.qmul.ac.uk[2]

ABSTRACT

This paper provides a method for comparing phase angles of harmonic sound sources. In particular, we propose an algorithm for decomposing the difference between two sets of phases into a harmonic part, which represents the phase progress of harmonic components, and a residue part, which represents all causes of deviations from perfect harmonicity. This decomposition allows us to compare phase alignments regardless of an arbitrary time shift, and handle harmonic and noise/inharmonic parts of the phase angle separately to improve existing algorithms that handles harmonic sound sources using phase measurements. These benefits are demonstrated with a new phase-based pitch marking algorithm and an improved time-scale and pitch modification scheme using traditional harmonic sinusoidal modelling.

1. INTRODUCTION

Pitched, or harmonic, sound sources produce periodical waveforms that can be represented as the sum of time-varying sinusoids (*partials*) whose frequencies are multiples of a fundamental *F*0[1][2]. This sinusoidal representation appears in audio processing either explicitly, e.g. in sinusoidal modelling[3], or implicitly, e.g. in the phase vocoder[4]. At any point each partial is associated with a *phase angle* that determines its positioning in time.

While each phase angle hardly makes any audible difference by itself, the alignment between phases affects the audio quality in various ways. Phase alignment between harmonic partials effects audible sub-period energy distribution [5], while that between binaural channels helps establish the perceived direction of the sound source [6]. A third type of phase alignment. i.e. that between phase angles sampled from the same partial at different time instants, affects sound quality via the frequency-phase relation. In particular, the perception of harmonicity relies on the partial frequencies being perfect multiples of a fundamental, which in turn requires the phase angles at different instants be aligned in a special way. An audio processing routine that does not preserve such phase alignment breaks the harmonicity, and possibly creates a chorus effect.

In this paper we address this third type of phase alignment in the context of pitched sounds. In particular, we compute a decomposition of the difference between two sets of phase angles into a harmonic progression and a least-square residue, the former evaluating how much progress the signal has made from one set of phases to the next, the latter evaluating whether the two sets are harmonically aligned, and if not, how far they are from being so. While being surprisingly simple, this treatment is useful in a variety of applications involving the handling of pitched sound sources.

2. HARMONIC PHASE ALIGNMENT AND MISALIGNMENT

Consider a set of M harmonically related sinusoids

$$s_m(t) = a_m \cdot \cos \varphi_m(t),\ t \in \mathbf{R},\ m = 1, \ldots, M, \tag{1a}$$

where

$$\varphi_m(t) = \varphi_m^0 + 2\pi m f_0 t,\ \forall m \tag{1b}$$

is the phase angle of the m^{th} harmonic *partial*. We define

$$\boldsymbol{\varphi}(t) = (\varphi_1, \ldots, \varphi_m)^{\mathsf{T}},\ \boldsymbol{\varphi}^0 = (\varphi_1^0, \ldots, \varphi_m^0)^{\mathsf{T}},\ \mathbf{m} = (1, \ldots, M)^{\mathsf{T}}. \tag{2}$$

It is trivial to show that

$$\boldsymbol{\varphi}(t) - \boldsymbol{\varphi}^0 = 2\pi f_0 t\, \mathbf{m}, \tag{3}$$

in which $f_0 t$ counts the number of periods between 0 and t. Phase values we encounter in actual computations are often subject to arbitrary modulo-2π shifts, so it's better to write (3) as

$$\boldsymbol{\varphi}(t) - \boldsymbol{\varphi}^0 = 2\pi f_0 t \cdot \mathbf{m} + 2\mathbf{k}\pi,\ \mathbf{k} \in \mathbf{Z}^M \tag{4a}$$

or

$$\boldsymbol{\varphi}(t) - \boldsymbol{\varphi}^0 \equiv 2\pi f_0 t \cdot \mathbf{m} \pmod{2\pi}. \tag{4b}$$

We say two vectors of phases $\boldsymbol{\varphi}_1 \in \mathbf{R}^M$ and $\boldsymbol{\varphi}_2 \in \mathbf{R}^M$ are *harmonically aligned* (*aligned* for short) if $\exists \delta \in \mathbf{R}$, so that

$$\boldsymbol{\varphi}_2 - \boldsymbol{\varphi}_1 \equiv \delta \cdot \mathbf{m} \pmod{2\pi}. \tag{5}$$

By this definition, the phase vector $\boldsymbol{\varphi}(t)$ in (4b) sampled at any time t is aligned to the initial $\boldsymbol{\varphi}^0$, while those sampled at any pair of t_1, t_2 are aligned between themselves. For various reasons, in real-world tasks the phase angles associated with harmonic sinusoids may not always satisfy (5), but carry an error term $\boldsymbol{\varepsilon}$:

$$\boldsymbol{\varphi}_2 - \boldsymbol{\varphi}_1 \equiv \delta \cdot \mathbf{m} + \boldsymbol{\varepsilon} \pmod{2\pi}. \tag{6}$$

We say $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ in (6) are *harmonically misaligned* (*misaligned* for short) by $\boldsymbol{\varepsilon}$. The following statements are equivalent:

i. $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ are harmonically misaligned by $\boldsymbol{\varepsilon}$;
ii. $\boldsymbol{\varphi}_2$ is harmonically aligned to $\boldsymbol{\varphi}_1 + \boldsymbol{\varepsilon}$;
iii. $\boldsymbol{\varphi}_2 - \boldsymbol{\varphi}_1$ is harmonically aligned to $\boldsymbol{\varepsilon}$;
iv. $\exists \delta \in \mathbf{R}$, $\mathbf{k} \in \mathbf{Z}^M$, so that

$$\boldsymbol{\varepsilon} = \boldsymbol{\varphi}_2 - \boldsymbol{\varphi}_1 - \delta \cdot \mathbf{m} + 2\mathbf{k}\pi. \tag{7}$$

Statement iv shows that (6) does not uniquely quantify $\boldsymbol{\varepsilon}$: any pair of *misalignments* $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are equivalent as long as they are aligned to each other. To compare phase alignments quantitatively, we'd like to quantify $\boldsymbol{\varepsilon}$ so that smaller $\boldsymbol{\varepsilon}$ is associated with phase vectors closer to perfect alignment. Particularly, if $\boldsymbol{\varphi}_1$ and

$\boldsymbol{\varphi}_2$ are harmonically aligned then $\boldsymbol{\varepsilon}$ should be $\mathbf{0}$. This leads to the minimum misalignment detailed below.

3. MINIMUM HARMONIC PHASE MISALIGNMENT

Equation (7) gives the general form of misalignment between $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$. The minimum misalignment method seeks to minimize $\boldsymbol{\varepsilon}$, constrained by (7), as the unique quantification of $\boldsymbol{\varepsilon}$. Different forms of the minimum can be defined by specific choices of the minimization criterion. In this paper we consider two of them, based on L^2 and weighted L^2 norms, respectively.

3.1. Minimization by L^2

The L^2 norm of $\boldsymbol{\varepsilon}$ is

$$\|\boldsymbol{\varepsilon}\|^2 = \boldsymbol{\varepsilon}^{\mathsf{T}}\boldsymbol{\varepsilon}. \tag{8}$$

Define $\boldsymbol{\varepsilon}^o = \boldsymbol{\varphi}_2 - \boldsymbol{\varphi}_1$, we can rewrite (7) as

$$\boldsymbol{\varepsilon}(\delta, \mathbf{k}) = \boldsymbol{\varepsilon}^o - \delta\cdot\mathbf{m} + 2\mathbf{k}\pi. \tag{9}$$

Once δ is fixed, the minimization of $\|\boldsymbol{\varepsilon}\|^2$ with regard to $\mathbf{k}$ is trivial: we only need to take the minimal absolute residue of $\boldsymbol{\varepsilon}^o - \delta\cdot\mathbf{m}$ modulo 2π:

$$\min_{\mathbf{k}} \boldsymbol{\varepsilon}(\delta, \mathbf{k}) = \operatorname{res}(\boldsymbol{\varepsilon}^o - \delta\cdot\mathbf{m}, 2\pi), \tag{10}$$

where res($\mathbf{x}$, y) is the minimal absolute residue of $\mathbf{x}$ modulo y, obtained by shifting every entry of $\mathbf{x}$ by a multiple of y into the interval [-y/2, y/2). Since $\boldsymbol{\varepsilon}$ is periodical with regard to δ with period 2π, the task of minimizing $\|\boldsymbol{\varepsilon}\|^2$ is simplified to finding $\delta\in[-\pi, \pi)$ so that

$$\|\boldsymbol{\varepsilon}(\delta)\|^2 = \sum_{m=1}^{M} \operatorname{res}(\varepsilon_m^o - m\delta, 2\pi)^2 \tag{11}$$

becomes minimal, where ε_m^o is the m^{th} entry of $\boldsymbol{\varepsilon}^o$.

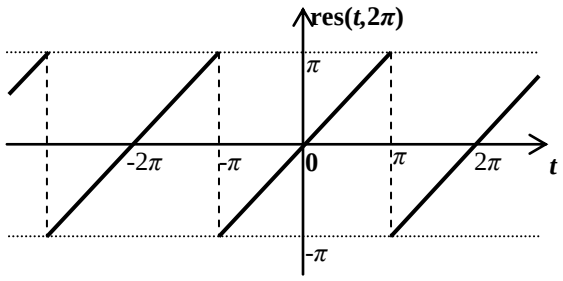


Figure 1 Minimal absolute residue modulo 2π

Since res(x,2π) is a piecewise linear function of x (Figure 1), $\|\boldsymbol{\varepsilon}(\delta)\|^2$ in (11) is a piecewise quadratic function of δ, whose minimum over the finite-length interval [$-\pi$, π) can be found by enumerating all quadratic pieces within this range.

3.2. Minimization by weighted L^2

L^2 minimization in 3.1 assumes equal impact from each partial. In practice it is often reasonable to emphasize some partials while deemphasizing some others. For example, some musical instruments have weaker even partials than odd ones, so that the phase angles measured for the odd partials are generally more reliable. It makes sense to emphasize the contribution from the stronger partials in formulating the minimization criterion.

The weighted L^2 norm of $\boldsymbol{\varepsilon}$ is

$$\|\boldsymbol{\varepsilon}\|_{\mathbf{w}}^2 = \boldsymbol{\varepsilon}^{\mathsf{T}}\cdot\operatorname{diag}(\mathbf{w})\cdot\boldsymbol{\varepsilon}. \tag{12}$$

where $\mathbf{w}=(w_1, \ldots, w_M)^{\mathsf{T}}$, $w_m\geq 0$, $\forall m$, contains the partial weights and diag($\mathbf{w}$) is a diagonal matrix whose main diagonal is specified by $\mathbf{w}$. $\|\boldsymbol{\varepsilon}\|_{\mathbf{w}}^2$ reduces to $\|\boldsymbol{\varepsilon}\|^2$ when all entries of $\mathbf{w}$ are 1.

The minimization of the weighted L^2 norm follows the same path as that of the L^2 norm. It is eventually reduced to finding $\delta\in[-\pi, \pi)$ that minimizes

$$\|\boldsymbol{\varepsilon}(\delta)\|_{\mathbf{w}}^2 = \sum_{m=1}^{M} w_m \cdot \operatorname{res}(\varepsilon_m^o - m\delta, 2\pi)^2 \tag{13}$$

A routine for the computation of δ in (11) is suggested in Appendix A, which also works for (11) if all weights are set to 1.

3.3. Interpretation

The minimum-by-L^2 method provides a least square solution to (6), which breaks the phase difference $\boldsymbol{\varphi}_2-\boldsymbol{\varphi}_1$ down to a harmonic progression $\delta\cdot\mathbf{m}$ and a least square mismatch term $\boldsymbol{\varepsilon}$.

If $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ are sampled from the same harmonic source, then δ estimates the phase progress (modulo 2π) of the fundamental frequency between the sampled positions, and $\boldsymbol{\varepsilon}$ evaluates their difference in phase sampling error. If $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ are sampled from different harmonic sources then δ estimates how much progress the first source has to make to be optimally aligned with the second source at the sampled position, and $\boldsymbol{\varepsilon}$ evaluates how well they can match in terms of phase alignments.

Although our discussion had started with steady tones, all formulations since (5) also apply to time-varying harmonic sound sources with amplitude and frequency modulations as well as timbre evolution. In the last case the phase vectors become landmarks during the transition from one timbre to the next, which can be used later to resynthesize the same transition.

Once we have estimated δ and $\boldsymbol{\varepsilon}$ we can write

$$\boldsymbol{\varphi}_2 - \boldsymbol{\varphi}_1 = (\delta + 2k\pi)\cdot\mathbf{m} + \boldsymbol{\varepsilon},\ k\in\mathbf{Z} \tag{14}$$

The difference from (7) is that the arbitrary factor is now $k\in\mathbf{Z}$ instead of $\mathbf{k}\in\mathbf{Z}^M$ in the original equation. This disambiguation comes from an implicit phase unwrapping during the minimization of $\boldsymbol{\varepsilon}$ with harmonicity constraint. This avoids possible loss of harmonicity during explicit phase unwrapping of each individual partial, e.g. in conventional sinusoidal synthesis [3].

4. APPLICATIONS IN PROCESSING HARMONIC SOUND SOURCES

The main advantage of our handling of the phase difference is that we may now attribute all pitch-related information to the harmonic progression part and focus on the residue part for handling non-harmonic aspects. By treating these two parts separately we can avoid the negative influence brought by one part to algorithms designed to handle the other.

In this section we demonstrate, in two unrelated applications, how we make use of the decomposition to handle harmonic sound sources. In a pitch marking example, we use the harmonic part to clock specific time instants within a period, and the residue part to inform on the degree of harmonicity. In another time-sale modification example, we demonstrate how we eliminate phase dispersion artefacts by avoiding frequency pollution from the misalignment part, while perfectly preserving wave shape evolution embedded in the misalignment.

4.1. Pitch marking

Pitch-synchronized processing, such as pitch-synchronous overlap-add (PSOLA)[7], operates on pitched sounds on the period level, offering quick response to period-to-period changes that is common in non-stationary quasi-harmonic signals like human speech. These algorithms rely on pre-determined time instants, known as *pitch marks*, as landmarks to synchronize their operations to. The basic requirement for pitch marking is that the interval between two adjacent pitch marks be precisely one speech period. Exactly where within a period the pitch mark should be placed has been a matter of the designer's choice: positions like prominent waveform peaks [8], significant excitation instants [9] and glottal closure instants [10] have all been reported for pitch marks. The chief concepts behind these choices are one: to clock the pitch marks consistently across periods and, if possible, across different sounds. The phase angle, by nature, handles exactly the clocking of an instant within the duration of a period. This motivates the new pitch marking algorithm presented below.

The standard procedure of pitch marking includes an initialization step, which places the first pitch mark, and a propagation step, which iteratively "grows" the existing set of pitch marks forward and backward to cover the whole length of a pitched sound. We will present our phase-based solution to the two steps in reverse order.

First let us define the phase vectors used in our algorithm. Let $s(t)$ be a periodical signal with period T_1. Given any time t_1, we consider the two-period interval $(t_1-T_1,\ t_1+T_1)$. A pitch-synchronized spectrum can be computed from the interval with

$$\hat{s}(k;t_1) = \sum_{t=t_1-T_1}^{t_1+T_1} w(t/T_1)s(t)e^{-j2\pi kt/2T_1},\ k = 0, 1, \ldots, T_1 \quad (15)$$

where k is the harmonics index and $w(t)$ is an analysis window supported on [-1, 1]. The phase vector of $s(t)$ of size M ($M<T_1$) at t_1 is then taken as

$$\boldsymbol{\varphi}_1 = \left(\arg\hat{s}(1;t_1), \arg\hat{s}(2;t_1), \cdots, \arg\hat{s}(M;t_1)\right)^{\mathrm{T}} \quad (16)$$

Now we let t_1 be where we have placed a pitch mark, and consider where to place the next. Ideally at point t_1+T_1 we should be able to sample the same phase vector $\boldsymbol{\varphi}_1$, which is rationale enough to place the next pitch mark there. However, in real tasks the signal is hardly exactly periodic, the estimate of T_1 is rarely perfect, and the period itself may have changed from T_1. Consequently at t_1+T_1 we only get some $\boldsymbol{\varphi}_2 \neq \boldsymbol{\varphi}_1$ which is not enough to signal the next pitch mark.

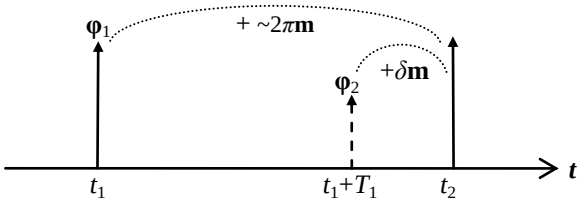


Figure 2 Pitch mark propagation by phase alignment

However, we can still place the next pitch mark *near* t_1+T_1 by harmonically progress $\boldsymbol{\varphi}_2$ by some $\delta\cdot\mathbf{m}$ to optimally match $\boldsymbol{\varphi}_1$. In other words, we put $\boldsymbol{\varphi}_1-\boldsymbol{\varphi}_2$ on the left side of (6) and solve for δ using the proposed routine (Appendix A). The new pitch mark is then placed at the adjusted position

$$t_2 \leftarrow t_1 + T_2,\quad T_2 = \frac{T_1}{1-\delta/2\pi}. \quad (17)$$

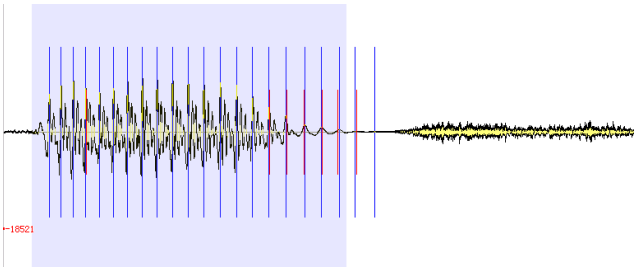
In (17) T_1 is multiplied by $2\pi/(2\pi-\delta)$, the ratio between the fundamental phase progression expected over a period and that observed over the duration of T_1. If the adjustment is large (i.e. $|\delta|$ is above some threshold), it makes sense to repeat the above adjustment process until δ is contained. Once the position of t_2 is determined, pitch marking may proceed from t_2 onwards with the new period T_2, until some termination criterion is met, such as the phase mismatch $\boldsymbol{\varepsilon}$ between pitch marks getting too large, or the correlation between marked periods getting too small. Backward propagation of pitch marks can be handled in exactly the same way.

We observe that the initial pitch mark provides a phase vector to which all other pitch marks are aligned, so its choice more or less determines what most pitch marks will be like. To better prepare for later stages that will use pitch marks, we would like all initial pitch marks have a consistent look. In [8], [9] and [10], this took the shape of waveform peak, excitation peak or glottal stop. For the phase-based pitch marking, we propose to initialize the first pitch mark at a position that is optimally aligned to the zero phase vector $\mathbf{0}^M$. This keeps the initial pitch mark close to minimum phase, which is related to high energy concentration and smooth spectral envelope preservation in PSOLA.

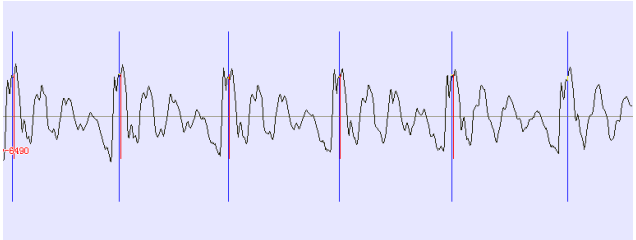
Let t_1 be a point around which the signal $s(t)$ has its period T_1 estimated with high periodicity, e.g. via autocorrelation. Let $\boldsymbol{\varphi}_1$ be the phase vector at t_1 given by (15)(16). We find out the harmonic phase progression $\delta\cdot\mathbf{m}$ for $\boldsymbol{\varphi}_1+\delta\cdot\mathbf{m}$ to optimally match $\mathbf{0}$. This is achieved by setting $\boldsymbol{\varphi}_2=\mathbf{0}$ on the left side of (6) and solve for δ with the proposed routine (Appendix A). We then update t_1 with

$$t_1 \leftarrow t_1 + T_1\cdot\delta/2\pi\ . \quad (18)$$

This update may be repeated a few times if it brings the match closer to zero phase, which can be observed from the value of $\|\boldsymbol{\varepsilon}\|^2$ before and after each update.



(a) waveform and pitch marks of "that's"



(b) detail of (a)

Figure 3 Pitch marking example 1

Figure 3(a) shows the pitch marking result for part of a spoken sentence in which a female speaker says "that's …". Pitch marks are plotted onto the waveform as vertical lines. As we may expect, the pitch marks span the duration of the voiced (i.e. periodic) part of the speech, leaving the fricative /s/ clear. Each period of

the signal contains exactly one pitch mark. An inspection into the details (Figure 3(b)) reveals that the pitch marks are placed at the same position within its period. Each pitch mark sits between two secondary waveform peaks riding on top of the highest of the four primary peaks of each period. This agrees with the common understanding of minimum phase, as well as that of the ideal position for the waveform grain centres in PSOLA.

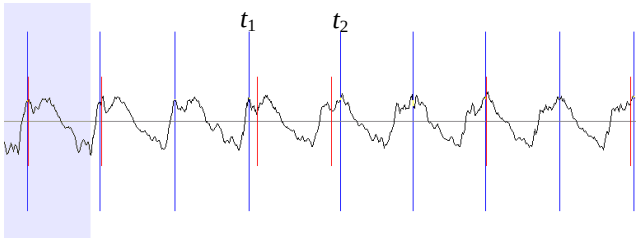


Figure 4 Pitch marking example 2

Figure 4 shows another pitch marking result during timbre evolution, in this case a change of phoneme without resting the vocal folds. We see that as the change progresses, the pitch mark shifts from one competing secondary peak (front peak at t_1) to another (rear peak at t_2). This is a common phenomenon with all landmark-based pitch markers: if each period contains two qualified landmarks, one slowly diminishes across the periods, and the other slowly rises, then a switch of the pitch mark between the two is sure to take place at least once during the transition. The advantage of our minimum phase pitch marking is that it also computes the fundamental phase progression from t_1 to t_2, so we are informed that t_2-t_1 exceeds a period's length, and by how much it does so. This information gains us better ground for adapting later processing stages, such as PSOLA, to handle such shifts properly.

4.2. Time-scale modification with sinusoidal model synthesis

Sinusoidal modelling [3] and its variants [2][11][12] represent sinusoids by sampling their amplitude, frequency and phase angle at predefined *measurement points* $\mathbf{t}=(t_0, \ldots, t_L)^\mathsf{T}$. In this part we consider only harmonic sinusoids. Let the angular frequency and phase angle of the m^{th} partial sampled at t_l be ω_l^m and φ_l^m, respectively, and let $\boldsymbol{\varphi}^m=(\varphi_0^m,\cdots,\varphi_L^m)^\mathsf{T}$, $\boldsymbol{\varphi}_l=(\varphi_l^1,\cdots,\varphi_l^M)^\mathsf{T}$, $\boldsymbol{\omega}^m=(\omega_0^m,\cdots,\omega_L^m)^\mathsf{T}$, $\boldsymbol{\omega}_l=(\omega_l^1,\cdots,\omega_l^M)^\mathsf{T}$. For each partial m, the sinusoidal model synthesis (SMS) reconstructs a new pair of frequency and phase functions $\tilde{\omega}^m(t)$ and $\tilde{\varphi}^m(t)$ by jointly interpolating $\boldsymbol{\varphi}^m$ and $\boldsymbol{\omega}^m$ with phase unwrapping, so that $\tilde{\omega}^m(t)$ is continuous and

$$\int_{t_l}^{t_{l+1}} \tilde{\omega}^m(t)dt \equiv \varphi_{l+1}^m-\varphi_l^m \pmod{2\pi},\ \forall l. \tag{19}$$

In the original SMS [3] $\tilde{\omega}^m(t)$ was constructed as piecewise quadratic. Finally the synthesizer reconstructs the phase with

$$\tilde{\varphi}^m(t)=\varphi_0^m+\int_{t_0}^{t}\tilde{\omega}^m(\tau)d\tau\,. \tag{20}$$

We consider the interval $[t_l, t_{l+1}]$, on which $\tilde{\omega}^m(t)$ depends on ω_l^m, ω_{l+1}^m, φ_l^m and φ_{l+1}^m. Let $\tilde{\boldsymbol{\omega}}_l(t)=\left(\tilde{\omega}^1(t),\cdots,\tilde{\omega}^M(t)\right)^\mathsf{T}$. We write $\tilde{\boldsymbol{\omega}}_l(t)=\tilde{\boldsymbol{\omega}}_l(t;\boldsymbol{\Omega},\boldsymbol{\varphi}_l,\boldsymbol{\varphi}_{l+1})$ to emphasize the dependency of $\tilde{\boldsymbol{\omega}}_l(t)$ on $\boldsymbol{\varphi}_l$ and $\boldsymbol{\varphi}_{l+1}$, with $\boldsymbol{\Omega}$ representing the frequencies. For all phase-aligned synthesizers, the dependency of $\tilde{\boldsymbol{\omega}}_l(t)$ on $\boldsymbol{\varphi}_l$ and $\boldsymbol{\varphi}_{l+1}$ is via their difference, so we can rewrite (19) in the following vector form:

$$\int_{t_l}^{t_{l+1}}\tilde{\boldsymbol{\omega}}_l(t;\boldsymbol{\Omega},\boldsymbol{\varphi}_{l+1}-\boldsymbol{\varphi}_l)dt\equiv\boldsymbol{\varphi}_{l+1}-\boldsymbol{\varphi}_l \pmod{2\pi}\,. \tag{21}$$

Ideally, for harmonic sinusoids the phase angles are always aligned, so that the right side of (21) is aligned to $\mathbf{0}$. In practice, due to inaccuracies of sampled frequencies and phases, $\boldsymbol{\varphi}(t_{l+1})$ is rarely perfectly aligned to $\boldsymbol{\varphi}(t_l)$, so that $\tilde{\boldsymbol{\omega}}(t)$ cannot be perfectly harmonic on $[t_l, t_{l+1}]$. This deviation from harmonicity is tolerable most of the time, as the misalignment of phase does not go much further beyond the magnitude of estimation error, so that the partials remain safe from destructive waveform interferences. However, this may not be the case when time-scale modification is involved.

The method

Sinusoidal modeling represents the time scale by the measurement points $\mathbf{t}$. Time-scale modification in this case involves selecting a new sequence $\mathbf{t}'$ and synthesizing a sound whose qualities (other than speed) at every t_l' are similar to those of the unmodified sound at t_l. In this paper we only consider the simplest type of time-scale modification, i.e. constant-rate time scaling, with

$$\mathbf{t}'=\rho\cdot\mathbf{t}\,. \tag{22}$$

where ρ is the scaling rate. The time scaling preserves the frequency value of $\tilde{\omega}^m(t)$ at point ρt, which we write as

$$\tilde{\boldsymbol{\omega}}_l'(\rho t)=\tilde{\boldsymbol{\omega}}_l(t;\boldsymbol{\Omega},,\boldsymbol{\varphi}_{l+1}-\boldsymbol{\varphi}_l)\,, \tag{23}$$

where $\tilde{\boldsymbol{\omega}}_l'(t)$ is the modified frequency function on $[t_l',t_{l+1}']$. Since the modified frequencies are linear stretching of the originals, the phase misalignment between $\boldsymbol{\varphi}_l$ and $\boldsymbol{\varphi}_{l+1}$ is multiplied by ρ. This change of misalignment can propagate across frames. As ρ becomes large the accumulated misalignment may incur destructive interference, which leads to deformed wave shape and eventually audible artefacts. This is known as *phase dispersion* and was well studied in [13].

From (23) we know that the phase progression of $\tilde{\boldsymbol{\omega}}_l'(t)$ between t_l' and t_{l+1}' is ρ times that of $\tilde{\boldsymbol{\omega}}(t)$ between t_l and t_{l+1}, including both the harmonic progression and the misalignment. However, for a harmonic sound source only the harmonic progression represents the true frequencies, which are what we aim to stretch by time scaling. The misalignment, on the other hand, represents estimation error and timbre evolution, neither a part of the true frequency, and should be left outside the integration of stretched frequency. As long as the misalignment is *not* multiplied by ρ, but remains unchanged between measurement points, there will be no extra misalignment to propagate across frame, therefore phase dispersion will not occur.

Back to synthesizer design, we break the phase difference $\boldsymbol{\varphi}_{l+1}-\boldsymbol{\varphi}_l$ into the sum of a harmonic progression and a least square residue:

$$\boldsymbol{\varphi}_{l+1}-\boldsymbol{\varphi}_l-2\mathbf{k}\pi=\delta\cdot\mathbf{m}+\boldsymbol{\varepsilon}\,,\ \mathbf{k}\in\mathrm{Z}^M. \tag{24}$$

Then we construct $\tilde{\boldsymbol{\omega}}_l(t)$ in two parts:

$$\begin{gathered}\tilde{\boldsymbol{\omega}}_l(t;\boldsymbol{\Omega},\boldsymbol{\varphi}_{l+1}-\boldsymbol{\varphi}_l)=\tilde{\boldsymbol{\omega}}_l^{//}(t)+\tilde{\boldsymbol{\omega}}_l^{\perp}(t),\\ \tilde{\boldsymbol{\omega}}_l^{//}(t)=\tilde{\boldsymbol{\omega}}_l(t;\boldsymbol{\Omega},\delta\cdot\mathbf{m}),\quad \tilde{\boldsymbol{\omega}}_l^{\perp}(t)=\tilde{\boldsymbol{\omega}}_l(t;\mathbf{0},\boldsymbol{\varepsilon})\end{gathered}\,. \tag{25}$$

$\tilde{\boldsymbol{\omega}}_l^{//}(t)$ represents the harmonic frequencies which should be stretched; $\tilde{\boldsymbol{\omega}}_l^{\perp}(t)$ represents the residue frequency whose *contribution to phase* should be stretched. These two different types of stretching are implemented in a simple combined form:

$$\tilde{\boldsymbol{\omega}}_l'(\rho t) = \tilde{\boldsymbol{\omega}}_l^{//}(t) + \rho^{-1}\tilde{\boldsymbol{\omega}}_{//}^{\perp}(t), \tag{26}$$

which replaces (23) in time scaling. An intuitive interpretation of (26) is that we multiply the residue $\tilde{\boldsymbol{\omega}}_l^{\perp}(t)$ by ρ^{-1} to counteract the ρ times stretching, so that its integration between measurement points remain unchanged. A similar solution exists for handling phase in constant rate pitch scaling, which faces the same phase dispersion issue.

Example 1: artificially synthesized sound

We illustrate the time scaling algorithm above using synthesized harmonic sinusoids with M=3, L=3. We place the measurement points at 0, T, 2T and 3T, T=128.The fundamental frequency is constant at 0.005. The frequency and phase values at the measurement points are given to the synthesizers with simulated errors: (0.15,0.15,-0.15,-0.15)/T for the fundamental frequency and (0.1,0.1,-0.1,-0.1)π for the fundamental phase. For the 2nd and 3rd partials the errors are rotated by 1 and 2 slots, respectively. We use scaling rates ρ=1, 2 and 4, and compare the results with a baseline synthesizer that runs the same workflow but without the proposed phase handling, and Ninness and Henriksen's method [13], which implemented "phase invariant" time scaling to address the phase dispersion issue.

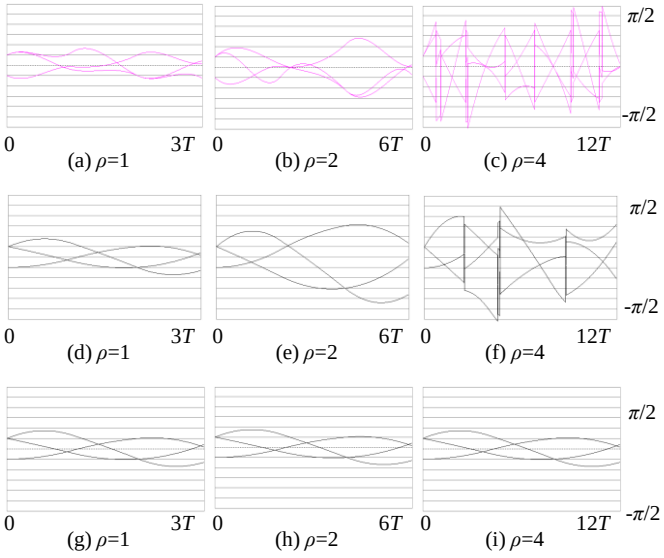


Figure 5 Phase misalignment during time scaling (a)(b)(c)Ninness-Henriksen method; (d)(e)(f)baseline method; (g)(h)(i)baseline+proposed.

Figure 5 shows the harmonic misalignment in each setting, computed between the error-free initial phases and the synthesized phases. Each curve in Figure 5 shows the misalignment of one sinusoidal partial against time (x-axis). Results from the Ninness-Henriksen method are given in the first row, those from the baseline in the second, those from the baseline with proposed phase handling in the third. We see that the method of [13] does offer smaller misalignment than the baseline at ρ=2, but both degrades to similar level at ρ=4. The performance of the proposed method, on the other hand, is not affected by time scaling.

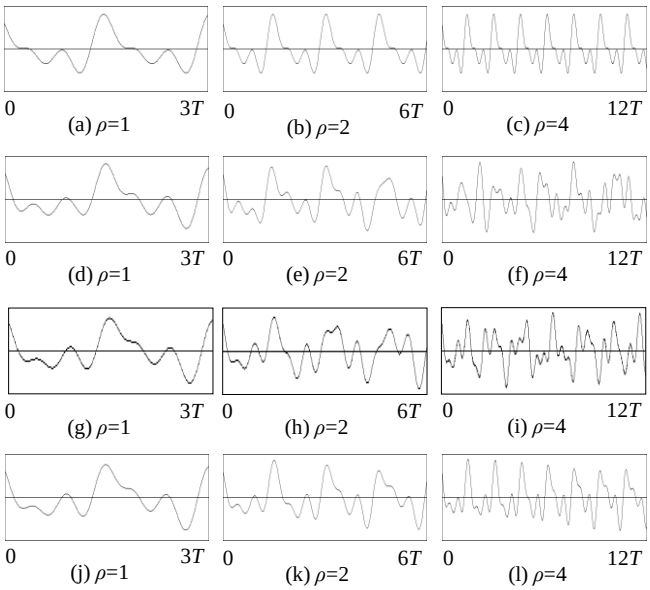
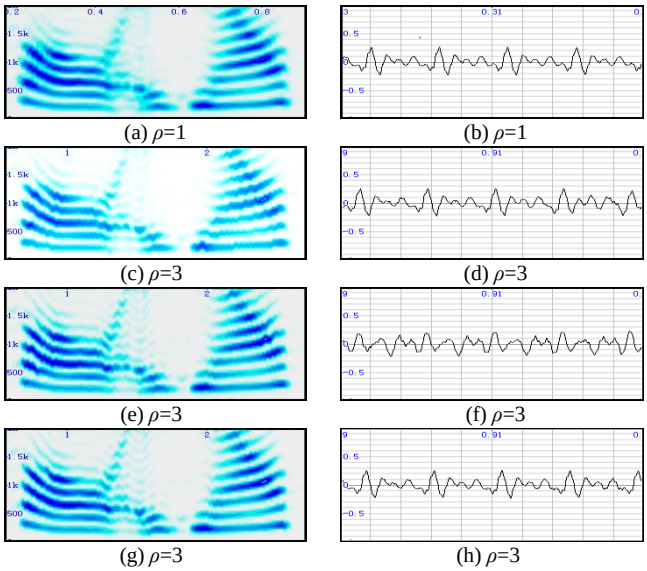


Figure 6 Waveforms before and after scaling (a)(b)(c)natural extension; (d)(e)(f)Ninness-Henriksen method; (g)(h)(i)baseline method; (j)(k)(l)baseline+proposed.

Figure 6 shows the synthesized waveforms. The partial amplitudes are assigned the ratio $1:2^{-1/2}:3^{-1/3}$, and the initial phases of the three partials are 0, π/7 and 4π/7. The first row gives the time scaling result obtained by natural extension of the ground truth signal; the second to fourth rows give the results from the phase invariant method of [13], the baseline synthesizer, and the baseline with proposed phase handling, respectively, using accurate amplitude values and inaccurate frequency and phase values at the measurement points. We see that Ninness and Henriksen's phase invariant method has succeeded to preserve better wave shape than our baseline at ρ=2, but at ρ=4 both lose hold of the waveform. The proposed method, on the other hand, preserves the waveform equally well for ρ=1, 2 and 4.

Example 2: voiced speech

In this example we time-stretch a recording of a female voice saying "Offal is now thought to be very nutritious." Using simple harmonic sinusoidal modelling without the residue, we extract the voiced part whose spectrogram and waveform are shown in Figure 7(a) and Figure 7(b). To visualize the details Figure 7(a) only contains the "-fal is now" part and Figure 7(b) only a few periods. Three time stretchers are applied to the sentence with ρ=3, including a plain phase-aligned synthesizer, the time stretcher proposed by Ninness and Henrikson [13], and our proposed approach. Their results are given in Figure 7(c) ~ Figure 7(h), aligned to their corresponding parts in the original in Figure 7(a) and Figure 7(b). The plain synthesizer combats phase dispersion by enforcing phase alignment at measurement points, at the cost of frequency instability. Ninness and Henrikson's approach smoothes out most frequency problems of the former but, judged from by the change in wave shape, has not managed to avoid phase dispersion. Our proposed time stretcher based on harmonic phase decomposition produces no less smooth result while perfectly maintains the wave shape.



(a) ρ=1 (b) ρ=1

(c) ρ=3 (d) ρ=3

(e) ρ=3 (f) ρ=3

(g) ρ=3 (h) ρ=3

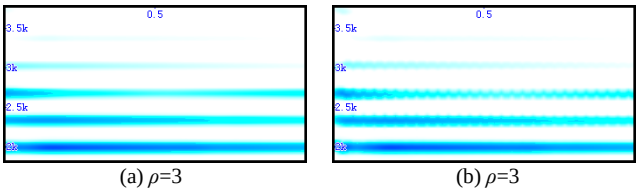
Figure 7 Spectrograms and waveforms (partial) before and after time scaling
(a)(b) original; (c)(d) plain phase-aligned synthesis;
(e)(f) Ninness-Henriksen; (g)(h) proposed.

Example 3: piano

In another experiment we time-stretch a piano note extracted from a polyphonic recording using harmonic sinusoidal modelling, using the same three stretchers as above with ρ=3. As far as subjective listening and wave shape comparison are concerned, our results for the piano note are very similar to those for the spoken sentence above. However, a closer look into spectrogram details in the frequency range above the first few reveals a vibrato-like structure in some partials synthesized using harmonic phase decomposition, which is absent from those synthesized using the Ninness-Henriksen method (Figure 8). Our informal listening has not detected audible traces of this frequency modulation, probably because the modulation has not affected the strongest partials sufficiently.

We attribute the modulation to the fact that the piano sound is intrinsically inharmonic, so the harmonic phase progression does not accurately model the relationship between its partial frequencies. When the phase angles are manipulated during time scaling using (26), the computed phase progression between measurement points may deviate significantly from what would be made at the actual frequency. The synthesizer makes up for the discrepancy by bending the instantaneous frequency between adjacent measurement points, creating a frequency modulation with period ρT.

More detailed analysis (using very large DFT size) of the average partial frequencies in Figure 8(b) reveals that the frequency values maintain harmonic ratios between themselves *in groups*. For example, the frequencies of the 8th, 7th and 6th partials have the ratio 8:7:6, those of the 10th and 9th partials have 10:9, while between the 9th and 8th the ratio is larger than 9:8. In other words, the harmonic rule may predict the average frequency of some partial from the immediate predecessor with a significant gap. This gap turns out to cover whole frequency bins, which is the result of phase unwrapping during sinusoidal synthesis.



(a) ρ=3 (b) ρ=3

Figure 8 Spectrogram details of time-stretched piano note
(a) Ninness-Henriksen method; (b) proposed.

5. CONCLUSION AND FUTURE WORK

In this paper we proposed comparing two sets of harmonic phase angles by decomposing their difference into a harmonic progression and a least square residue, computed by minimizing a piecewise binomial function. The method does not mind how the phase angles are computed, nor requires the phase angle of every partial be available. One is allowed to attach weights to the phase values during the decomposition to suppress the contribution from the inaccurate phases, particularly those of weak partials.

Compared to raw phase values, this harmonically decomposed representation has several advantages: it allows comparison of phase vectors regardless of an arbitrary time shift; it allows clocking the difference between two phase vectors using information from all partials; it allows algorithms designed for processing periodic signals to focus on harmonic phase, those for non-periodic aspects on the residue. In two unrelated applications involving harmonic sound sources, we have demonstrated the use of our analysis approach to phase in two different manners, both achieving expected results. However, since phase alignment is such a common presence with harmonic and quasi-harmonic sounds, the proposed method is surely applicable in many more circumstances. For example, we have already demonstrated in Figure 5 the use of harmonic misalignment to track timbre change, which is probably more revealing than comparing waveforms, such as Figure 6.

On the other hand, we have seen that the harmonic phase decomposition comes with a few limitations. First, it requires harmonic sinusoidal analysis to provide reasonably accurate phase values of harmonically related partials, which can be a hard task in itself, especially in complex acoustical environments. Whether and how the proposed technique may be applied to more readily available forms of phase angles, such as that from the Fourier transform, remains a question to be looked into. Second, successful decomposition of the phase progression relies on good phase estimates of all participating partials, which in turn requires a mechanism to tag each phase estimate with a confidence label. In this paper we have included partial weights in section 3.2 to fill this role, but how the weights are to be best evaluated remains another question for future investigation. Moreover, the piano example shows that the proposed method does not accurately model sound sources with inharmonicity. In the case of time scaling, this has lead to additional frequency modulation of some partials, and has the potential to create audible artefacts. The adaptation of the proposed harmonic phase decomposition to sound sources with inharmonicity, therefore, may become another direction of future research into this world of aligned and misaligned phase angles.

6. ACKNOWLEDGMENTS

Part of this work was completed at Queen Mary, University of London.

7. REFERENCES

[1] D. Erro, I. Sainz, E. Navas and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol.8 no.2, April 2014, pp.184-194.

[2] Wen X. and M. Sandler, "Sinusoid modeling in a harmonic context," in *Proceedings of DAFx'07*, Bordeaux, 2007.

[3] R.J. McAulay and T.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.34, no.4, 1986, pp.744-754.

[4] R. Fischman, "The phase vocoder: theory and practice," *Organised Sound*, vol.2 no.2, July 1997, pp.127-145.

[5] H. Pobloth and W.B. Kleijn, "On phase perception in speech," in Proceedings of *ICASSP'99*, Phoenix, 1999.

[6] D.J. Breebaart, F. Nater and A.G. Kohlrausch, "Parametric binaural synthesis: Background, applications and standards,", in *NAG-DAGA 2009: International Conference on Acoustics*, Rotterdam, 2009.

[7] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication* 9 (1990), Elsevier, North-Holland, pp.453-467.

[8] A.Chalamandaris, P. Tsiakoulis and S. Karabetsos, "An efficient and robust pitch marking algorithm on the speech waveform for TD-PSOLA," in *Proc. ICSIPA'09*, Kuala Lumpur, 2009.

[9] P.S. Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," *IEEE Transactions on Speech and Audio Processing*, vol.7 no.6, November 1999, pp.609-619.

[10] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. InterSpeech 2009*, Brighton, 2009.

[11] X. Serra and J. Smith III, "Spectral modeling synthesis: a sound analysis/synthesis based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol.14, 1990.

[12] S. N. Levine and J. Smith III, "A sines+transients+noise audio representation for data compression and time/pitch scale modifications," in *Proc. AES 105th Convention*, San Francisco, 1998.

[13] B. Ninness and S.J. Henriksen, "Time-scale modification of speech signals," *IEEE Transactions on Signal Processing*, vol.56 no.44, 2008.

APPENDIX A: COMPUTING MINIMUM HARMONIC PHASE MISALIGNMENT

Given $\boldsymbol{\varepsilon}^o=(\varepsilon_1^o, \ldots, \varepsilon_M^o)^\mathsf{T}$, $\mathbf{w}=(w_1, \ldots, w_M)^\mathsf{T}$, $w_m \geq 0$, $\forall m$, find $\delta \in [-\pi, \pi)$ that minimizes

$$D(\delta)=\sum_{m=1}^{M} w_m \cdot \mathrm{res}(\varepsilon_m^o + m\delta, 2\pi)^2 . \quad \text{(A.1)}$$

Notice that (A.1) gives δ a different sign from (13), which needs be switched back should the value of δ be required afterwards.

Because $\mathrm{res}(x,2\pi)$ is a piecewise linear regarding x, $D(\delta)$ is piecewise quadratic regarding δ. Let $\delta_0 = -\pi < \delta_1 < \ldots < \delta_L = \pi$ be the section points marking the end of these quadratic pieces, and let the left and right derivatives of $D(\delta)$ at δ_l be $D'(\delta_l)_-$ and $D'(\delta_l)_+$, then $D(\delta)$ has a local minimum in (δ_l, δ_{l+1}) if and only if $D'(\delta_l)_+ < 0$ and $D'(\delta_{l+1})_- > 0$; $D(\delta)$ has no minimum at any δ_l other than $-\pi$ and π as it has $-\infty$ derivative as such points. To find δ that minimizes $D(\delta)$, we first locate all the section points and compute the left and right derivatives, then enumerate the quadratic pieces for local minima, from which the smallest one is picked as the global minimum of $D(\delta)$.

Below is a routine for minimizing $D(\delta)$ that iteratively locates the section points contributed from each partial m and updates a section point list with derivatives.

routine 1: minimum harmonic phase misalignment

This routine maintains a list of section points $\{p_l, a_l, b_l\}$ indexed by l, in which p_l is the position of a section point and a_l, b_l are the left and right derivatives at p_l.

1 º Initialize a sorted section point list with initial members $-\pi$ and π, both with left and right derivatives set to 0;

2 ºfor m=1, …, M, do 3 ~8 º;

3 ºcompute the first section point of partial m:

$$sp_0 \leftarrow -\pi + (\pi - \mathrm{res}(-m\pi + \varepsilon_m^o)) / m; \quad \text{(A.2)}$$

4 ºfor $sp=sp_0$, $sp_0 + 2\pi/m$, …, $sp_0 + 2\pi(m-1)/m$, do 5 º;

5 ºif sp does not coincide with an existing point in the list, let the two listed section points immediately before and after sp be p_l and p_{l+1}, then we insert sp into the list with identical left and right derivatives given as

$$\frac{(p_{l+1}-sp)b_l + (sp-p_l)a_{l+1}}{p_{l+1}-p_l}; \quad \text{(A.3)}$$

6 ºfor all points p_l, l=0, 1, ..., in the updated list, do 7 ~8 º;

7 ºif p_l is a section point of partial m, do

$$a_l \leftarrow a_l + m\pi \cdot w_m, \quad \text{(A.4)}$$

$$b_l \leftarrow b_l - m\pi \cdot w_m; \quad \text{(A.5)}$$

8 ºif not, do

$$a_l \leftarrow a_l + m \cdot \mathrm{res}(mp_l + \varepsilon_m^o, 2\pi) \cdot w_m, \quad \text{(A.6)}$$

$$b_l \leftarrow b_l + m \cdot \mathrm{res}(mp_l + \varepsilon_m^o, 2\pi) \cdot w_m; \quad \text{(A.7)}$$

9 ºinitialize the minimum $\delta_{\min} \leftarrow -\pi$;

10 ºfor l=1, 2…, do 11 ~13 º;

11 ºif $b_{l-1}<0$ and $a_l>0$, do

12 ºcompute local minimum

$$\delta \leftarrow \frac{p_{l-1}a_l - p_l b_{l-1}}{a_l - b_{l-1}}; \quad \text{(A.8)}$$

13 ºif $D(\delta) < D(\delta_{\min})$, $\delta_{\min} \leftarrow \delta$.