MagPhase vocoder: magnitude and phase analysis/synthesis for statistical parametric speech synthesis

THE UNIVERSITY of EDINBURGH

{Felipe Espic, Cassia Valentini-Botinhao, Simon King }

The Centre for Speech Technology Research (CSTR)
University of Edinburgh, UK

Summary

We propose a simple new representation for the FFT spectrum and synthesis method tailored to statistical parametric speech synthesis. It consists of four feature streams that describe magnitude, phase, and fundamental frequency. The proposed feature extraction method does not decompose the speech structure (e.g., into source+filter or harmonics+noise). By avoiding these decompositions and using phase information, we can dramatically reduce the "phasiness" and "buzziness".
The method uses computationally cheap operations and can operate at a lower frame rate than the 200 frames-per-second typical in many systems. It avoids heuristics and methods requiring approximate or iterative solutions for phase unwrapping.
Subjective comparisons were made with a state-of-the-art baseline, using the STRAIGHT vocoder. In all variants, the proposed method substantially outperformed the baseline.

Motivation

- Vocoders have been identified as a cause of "buzziness" and "phasiness" [1].
- Source-filter model:
  - Source: Pulse train, white noise, glottal pulse modelling, or natural speech [2].
  - Filter: Smoothed spectral envelope with minimum-phase assumption.
- Sinusoidal modelling: Time-varying number of parameters.
- No phase modelling: Minimum-phase assumption, Griffin-Lim reconstruction.

Goals

The goals for the proposed approach are to:

- Minimise signal processing.
- Simplify estimation.
- Extract consistent features suitable for statistical modelling.
- Eliminate "phasiness" and "buzziness".
- Work with any standard real-valued deep learning method.

References

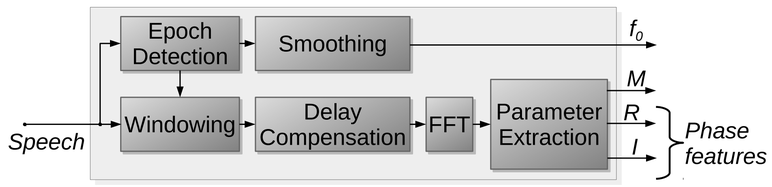
[1] T. Merritt, J. Latorre, and S. King, "Attributing modelling errors in HMM synthesis by stepping gradually from natural to modelled speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Brisbane), pp. 4220–4224, Apr. 2015.

[2] F. Espic, C. Valentini-Botinhao, Z. Wu, and S. King, "Waveform generation based on signal reshaping for statistical parametric speech synthesis," in *Proc. Interspeech*, (San Francisco, CA, USA), pp. 2263–2267, September 2016.

[3] F. Espic, C. Valentini-Botinhao, and S. King, "Direct modelling of magnitude and phase spectra for statistical parametric speech synthesis," in *Proc. Interspeech*, (Stochohlm, Sweden), August 2017.

Proposed Method

I. ANALYSIS:



Steps:

1. f_0 extraction by epoch detection.
2. Framing and windowing ($2T_0$ length).
3. Extraction of phase features (R, I):
   - Trick 1: Delay compensation.
   - Trick 2: Phase re-wrapping.

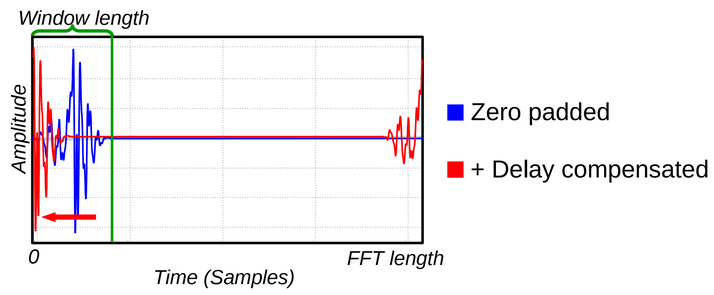


Figure 1: Delay compensation example.

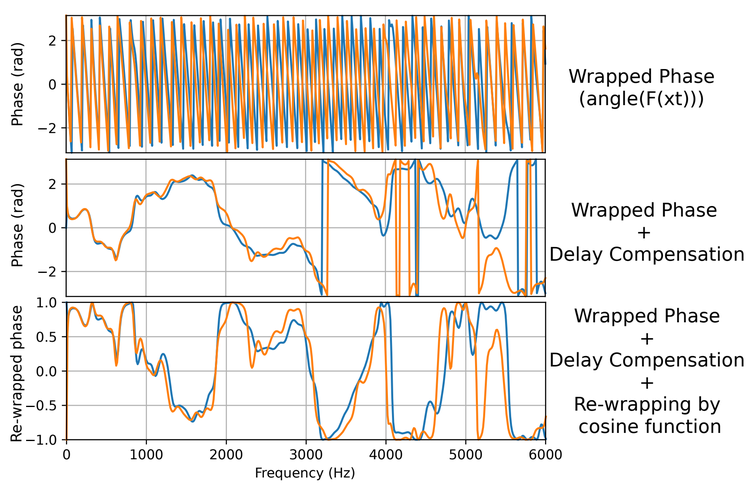


Figure 2: Example: extraction of feature R from two consecutive frames.

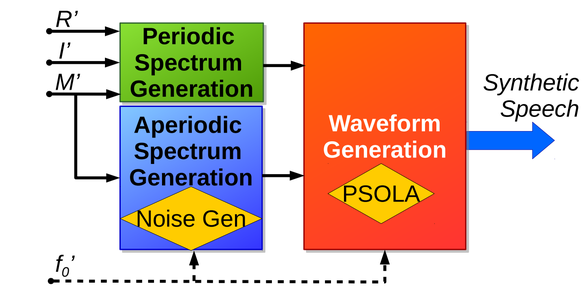
4. Magnitude spectrum computation.

EXTRACTED FEATURES:

$$\mathbf{M} = abs(\mathrm{X}_\omega)$$
$$\mathbf{R} = Real\{\mathrm{X}_\omega\}/abs(\mathrm{X}_\omega)$$
$$\mathbf{I} = Imag\{\mathrm{X}_\omega\}/abs(\mathrm{X}_\omega)$$
$$\boldsymbol{f0}_{[t]} = (e_{[t]} - e_{[t-1]})^{-1}$$

Where $\mathrm{X}_\omega = FFT(x_t)$

II. SYNTHESIS:



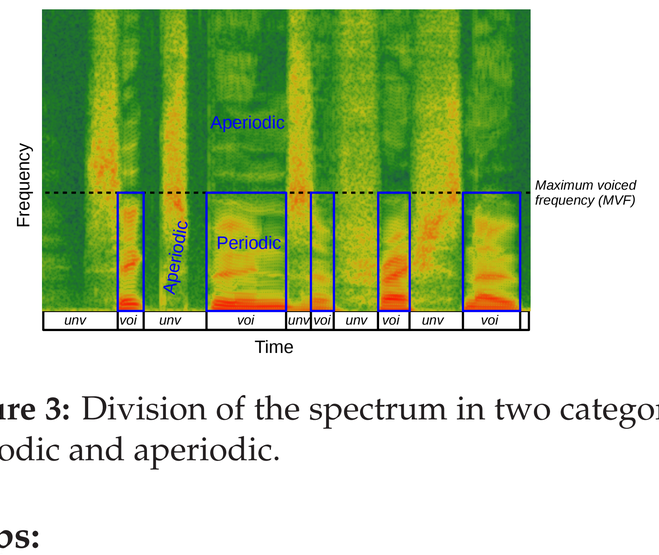


Figure 3: Division of the spectrum in two categories: periodic and aperiodic.

Steps:

1. Periodic Spectrum:
   $\mathrm{S}_{\mathrm{per}} = \mathrm{M}' \cdot (\mathrm{R}' + \mathrm{I}' \cdot j)/\sqrt{\mathrm{R}'^2 + \mathrm{I}'^2}$
2. Aperiodic Spectrum:
   $\mathrm{S}_{\mathrm{aper}} = \mathrm{M}' \cdot \mathcal{F}(x_{noise})$
3. Waveform Generation:
   $x_{syn} = \mathcal{F}^{-1}(merge(\mathrm{S}_{\mathrm{per}}, \mathrm{S}_{\mathrm{aper}})) \Rightarrow PSOLA$

Experiments

Two English text-to-speech voices (female and male) were built using the Merlin toolkit (SLSTM). The features extracted by the proposed method were compressed in frequency domain by Mel-scale:
$R = 45 \times Mel(R)$, $I = 45 \times Mel(I)$,
$lf0 = \log(f_0)$, $logM = 60 \times Mel(\log(M))$.
The systems were evaluated by 30 native English speakers using a MUSHRA-like test. Each subject evaluated 36 different sentences (18 male, 18 female). The systems under test were:

- **Nat:** Natural speech (the hidden reference).
- **Base:** The Baseline system running at constant frame rate and using STRAIGHT for analysis/synthesis.
- **PM:** The Proposed Method with settings as described in the paper [3].
- **PMv1:** The Proposed Method with voiced segments having **no** aperiodic component.
- **PMv2:** The Proposed Method with voiced segments having **no** aperiodicity window.

Speaker	Nat	Base	PM	PMv1	PMv2
Male	100	43.6	**51.4**	45.6	49.4
Female	100	32.6	**43.8**	34.6	43.1

Conclusions - Future Work

- New waveform analysis/synthesis method for SPSS.
- Does not require estimation of spectral envelope, aperiodicity, harmonics, etc.
- No iterative or estimation process beyond the epoch detection.
- Reduces "buzziness" and "phasiness".
- Outperforms state-of-the-art (STRAIGHT).
- No heuristics for phase modelling.
- Pitch synchronous processing throughout (no converesions to/from fixed rate), which can decrease frame rate by up to 31.5%
- Potential usability in other audio applications (e.g., ASR)
- Future work: Remove MVF, V/UV decision and $f0$ modelling.

Code and samples: `http://felipeespic.com/magphase/`