A Pulse Model in Log-domain for a Uniform Synthesizer

Gilles Degottex, Pierre Lanchantin, Mark Gales

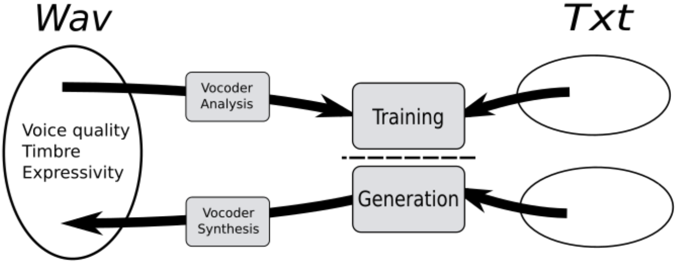
University of Cambridge - Engineering Department, UK

EU/Marie Sklodowska-Curie Fellowship, 2015-2017

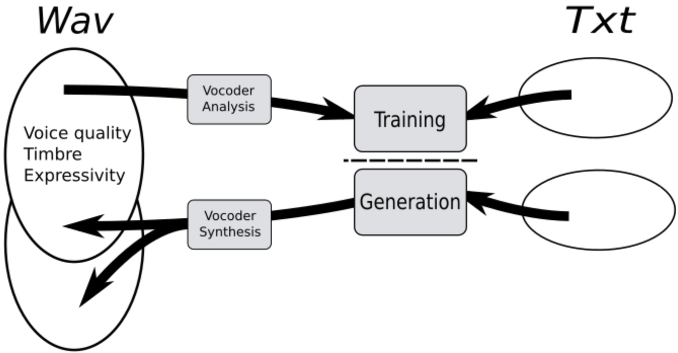
http://gillesdegottex.eu/Demos/HQSTS/

Motivation

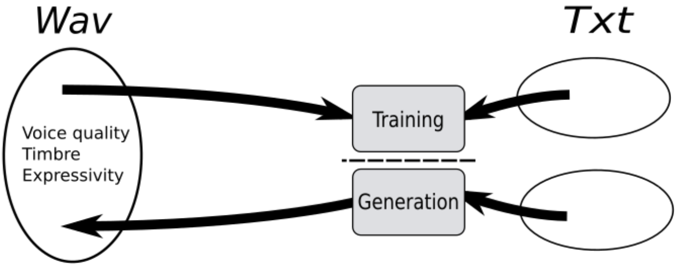
Traditional SPSS



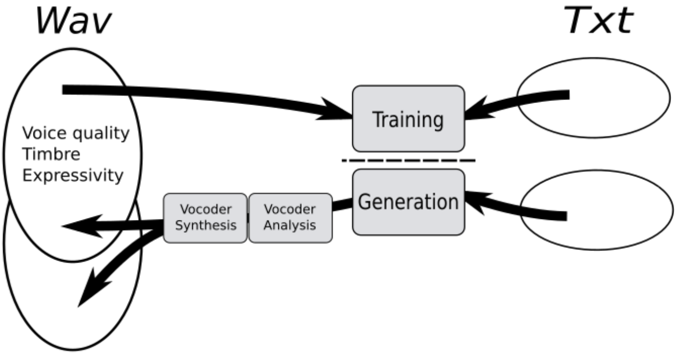
Traditional SPSS + Transformation



Direct waveform synthesis



Direct waveform synthesis + Transformation



Problem

Problem

Current Statistical Parametric Speech Synthesis (SPSS) quality is conditionned by the vocoder's quality.

Problem

Current Statistical Parametric Speech Synthesis (SPSS) quality is conditionned by the vocoder's quality.

Two approaches

Problem

Current Statistical Parametric Speech Synthesis (SPSS) quality is conditionned by the vocoder's quality.

Two approaches

A Drop the vocoder

Problem

Current Statistical Parametric Speech Synthesis (SPSS) quality is conditionned by the vocoder's quality.

Two approaches

A Drop the vocoder

B Do a better vocoder

Problem

Current Statistical Parametric Speech Synthesis (SPSS) quality is conditionned by the vocoder's quality.

Two approaches

A Drop the vocoder

B Do a better vocoder

In both cases, preferable to get closer to the waveform:

⇒ Simpler signal model

[A] might be embeded in the statistical model

[B] easier to implement, understand and control

⇒ More generic/uniform features

[A] might be used to bring perceptual a priori for training

[B] less constraints, absorb more the variations of the signal

Problem

Current Statistical Parametric Speech Synthesis (SPSS) quality is conditionned by the vocoder's quality.

Two approaches

A Drop the vocoder

B Do a better vocoder

In both cases, preferable to get closer to the waveform:

$\Rightarrow$ Simpler signal model

[A] might be embeded in the statistical model

[B] easier to implement, understand and control

$\Rightarrow$ More generic/uniform features

[A] might be used to bring perceptual a priori for training

[B] less constraints, absorb more the variations of the signal

Move the signal complexity to the features (good for machine learning!)

Some vocoder's pros and cons

STRAIGHT

+ Robust spectral envelope estimation
− Voicing decisions embedded into the synthesizer
− Sounds buzzy for high-pitched voices

Some vocoder's pros and cons

STRAIGHT

+ Robust spectral envelope estimation
– Voicing decisions embedded into the synthesizer
– Sounds buzzy for high-pitched voices

Harmonic Model + Phase Distortion (HMPD)

+ Cont. F0, uniform voiced/unvoiced repres. (voicing decision in the feature)
+ No buzziness for high-pitched voices (proper randomized mid-high freq.)
– Tensness in voiced segments (no noise between harmonics)

Some vocoder's pros and cons

STRAIGHT

- $+$ Robust spectral envelope estimation
- $-$ Voicing decisions embedded into the synthesizer
- $-$ Sounds buzzy for high-pitched voices

Harmonic Model + Phase Distortion (HMPD)

- $+$ Cont. F0, uniform voiced/unvoiced repres. (voicing decision in the feature)
- $+$ No buzziness for high-pitched voices (proper randomized mid-high freq.)
- $-$ Tenseness in voiced segments (no noise between harmonics)

Many others vocoders

- $-$ Complex (more prone to estim. errors $\Rightarrow$ artefacts)

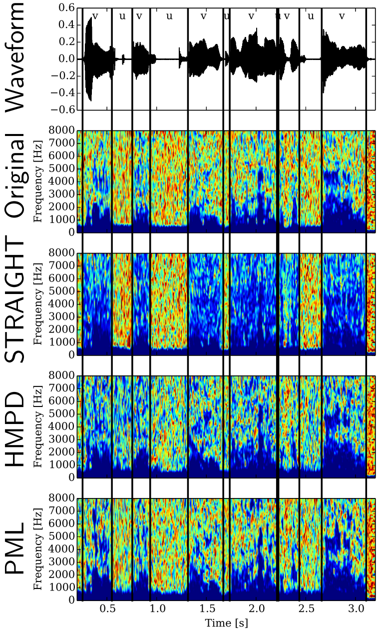
Vocoder's PDD Example

Phase Distortion Deviation (PDD) measures phase variance.

Run it on original and vocoders' resynthesis.

The warmer, the noisier.
The colder, the more deterministic.

Used for voice quality classification, voice pathology detection, vocoding.



Vocoder's PDD Example

Phase Distortion Deviation (PDD)

measures phase variance.

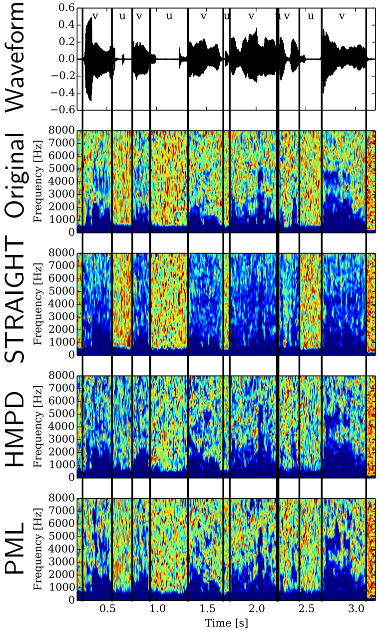
Run it on original and vocoders' resynthesis.

The warmer, the noisier.
The colder, the more deterministic.

Used for voice quality classification, voice pathology detection, vocoding.

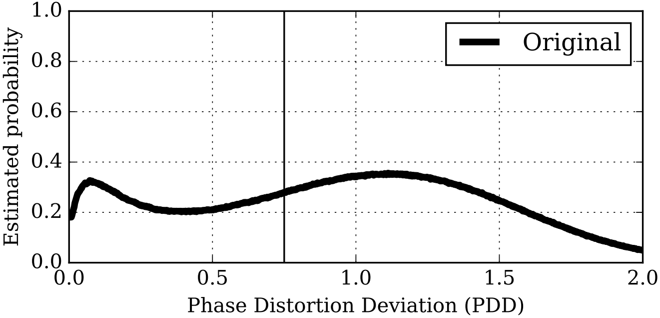
Observations

- Most vocoders fail to reconstruct the original.
- STRAIGHT exhibits very low noise in unvoiced segments.



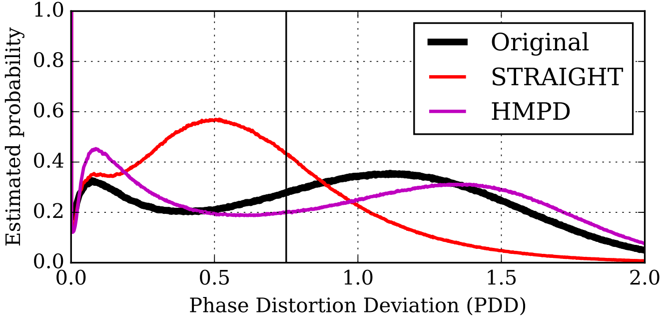
Vocoder's PDD Histograms

PDD histograms in voiced segments.
(of CMU SLT female voice)



Vocoder's PDD Histograms

PDD histograms in voiced segments
over analysis/resynthesis of 2 vocoders.



Idea

PML Signal model

Train of pulses $t_i = t_{i-1} + \frac{1}{f_0(t_{i-1})}$ (voiced and unvoiced!)

PML Signal model

Train of pulses $t_i = t_{i-1} + \frac{1}{f_0(t_{i-1})}$ (voiced and unvoiced!)

$$S_i(\omega) = e^{-j2\pi t_i} \cdot V(t_i, \omega) \cdot N_i(\omega)^{M(t_i,\omega)}$$

$e^{-j2\pi t_i}$ Time position

$V(t_i, \omega)$ Filter

$N_i(\omega)$ Gaussian noise in $[\frac{t_{i-1} - t_i}{2}, \frac{t_{i+1} - t_i}{2}]$

$M(t_i, \omega)$ Binary noise mask

PML Signal model

Train of pulses $t_i = t_{i-1} + \frac{1}{f_0(t_{i-1})}$ (voiced and unvoiced!)

$$S_i(\omega) = e^{-j2\pi t_i} \cdot V(t_i,\omega) \cdot N_i(\omega)^{M(t_i,\omega)}$$

$e^{-j2\pi t_i}$ Time position

$V(t_i,\omega)$ Filter

$N_i(\omega)$ Gaussian noise in $[\frac{t_{i-1}-t_i}{2}, \frac{t_{i+1}-t_i}{2}]$

$M(t_i,\omega)$ Binary noise mask

$$lS_i(\omega) = \overbrace{-j2\pi t_i}^{\text{Position}} + \overbrace{\log|V(t_i,\omega)|}^{\text{Amplitude}} + \overbrace{j\angle V(t_i,\omega)}^{\text{Minimum phase}}$$

PML Signal model

Train of pulses $t_i = t_{i-1} + \frac{1}{f_0(t_{i-1})}$ (voiced and unvoiced!)

$$S_i(\omega) = e^{-j2\pi t_i} \cdot V(t_i, \omega) \cdot N_i(\omega)^{M(t_i,\omega)}$$

$e^{-j2\pi t_i}$ Time position

$V(t_i, \omega)$ Filter

$N_i(\omega)$ Gaussian noise in $[\frac{t_{i-1}-t_i}{2}, \frac{t_{i+1}-t_i}{2}]$

$M(t_i, \omega)$ Binary noise mask

$$lS_i(\omega) = \overbrace{-j2\pi t_i}^{\text{Position}} + \overbrace{\log|V(t_i,\omega)|}^{\text{Amplitude}} + \overbrace{j\angle V(t_i,\omega)}^{\text{Minimum phase}}$$
$$+ \underbrace{M(t_i,\omega)}_{\text{Noise extent}} \cdot \underbrace{j\angle N_i(\omega)}_{\text{Phase randomi.}}$$

Phase randomization is great for removing buzziness!

PML Signal model

Train of pulses $t_i = t_{i-1} + \frac{1}{f_0(t_{i-1})}$ (voiced and unvoiced!)

$$S_i(\omega) = e^{-j2\pi t_i} \cdot V(t_i,\omega) \cdot N_i(\omega)^{M(t_i,\omega)}$$

- $e^{-j2\pi t_i}$ Time position
- $V(t_i,\omega)$ Filter
- $N_i(\omega)$ Gaussian noise in $[\frac{t_{i-1}-t_i}{2}, \frac{t_{i+1}-t_i}{2}]$
- $M(t_i,\omega)$ Binary noise mask

$$\begin{aligned} lS_i(\omega) = & \overbrace{-j2\pi t_i}^{\text{Position}} + \overbrace{\log|V(t_i,\omega)|}^{\text{Amplitude}} + \overbrace{j\angle V(t_i,\omega)}^{\text{Minimum phase}} \\ & + \underbrace{M(t_i,\omega)}_{\text{Noise extent}} \cdot \Big(\underbrace{j\angle N_i(\omega)}_{\text{Phase randomi.}} + \underbrace{\log|N_i(\omega)|}_{\text{Noise amplitude}} \Big) \end{aligned}$$

Phase randomization is great for removing buzziness!

PML Features

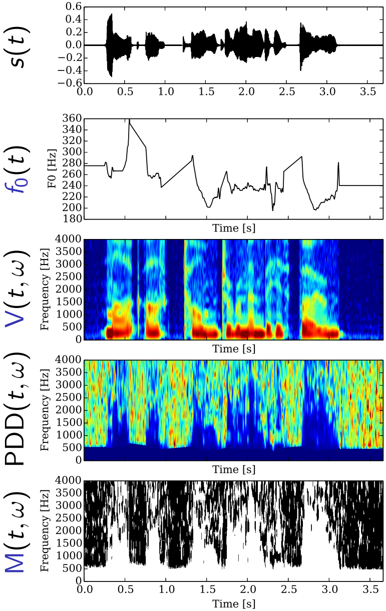
Train of pulses $t_i = t_{i-1} + \frac{1}{f_0(t_{i-1})}$

$$S_i(\omega) = e^{-j2\pi t_i} \cdot V(t_i, \omega) \cdot N_i(\omega)^{M(t_i,\omega)}$$

$f_0(t)$ Continuous fundamental frequency

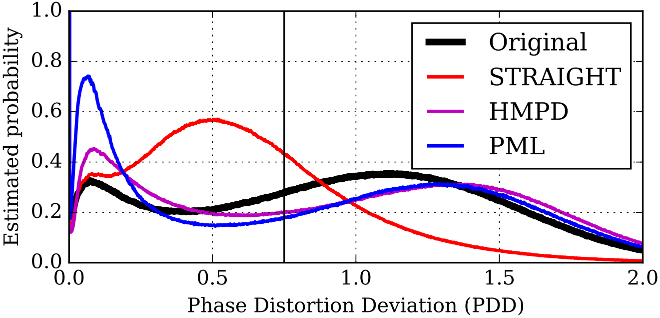
$V(t, \omega)$ Filter

$M(t, \omega)$ Binary noise mask (PDD thresh. at 0.75)



PML's PDD Histograms

PDD histograms in voiced segments
over analysis/resynthesis of 3 vocoders.

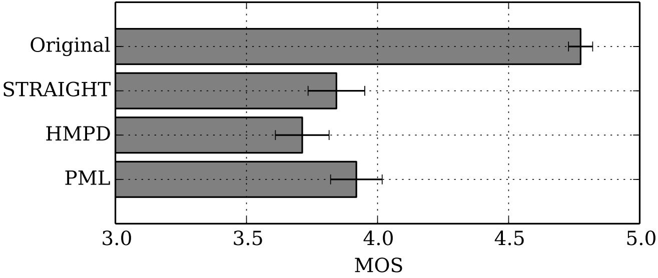


Experiments

Analysis/Resynthesis quality

Mean Opinion Scores (MOS) (with the 95% confidence intervals) of the analysis/resynthesis quality of 3 vocoders.

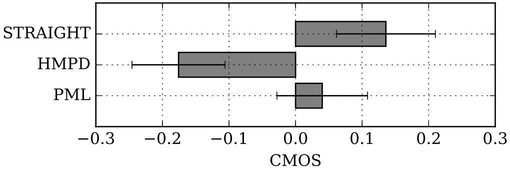
(6 voices: 4 American, 2 British; 3 females, 3 males)



Statistical Parametric Speech Synthesis (SPSS)

- 6 voices
  4 American, 2 British
  3 female, 3 male
- HTS for phonetic alignment
- HTS for duration model
- DNN for acoustic model:
  Input: 601 contextual input features
  Hidden: 6 × 1024 tanh
  Output for STRAIGHT:
    F0, VUV, MCEP, MAPER
  Output for HMPD and PML:
    F0, MCEP, MPDD

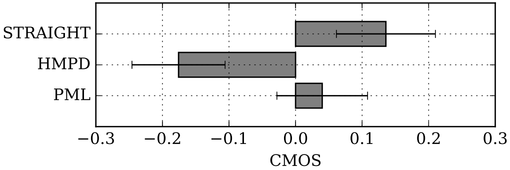
- Comparative Mean Opinion Score (CMOS)
  Listening test on AMTurk
- 53 Participants
  evaluated the 3 vocoders' combination of
  8 sentences among 142x6 sentences



Statistical Parametric Speech Synthesis (SPSS)

- 6 voices
  4 American, 2 British
  3 female, 3 male
- HTS for phonetic alignment
- HTS for duration model
- DNN for acoustic model:
  Input: 601 contextual input features
  Hidden: 6 × 1024 tanh
  Output for STRAIGHT:
    F0, VUV, MCEP, MAPER
  Output for HMPD and PML:
    F0, MCEP, MPDD

- Comparative Mean Opinion Score (CMOS)
  Listening test on AMTurk
- 53 Participants
  evaluated the 3 vocoders' combination of 8 sentences among 142x6 sentences



Conclusions

- PML solves major drawbacks in HMPD
  (while still using continuous f0 and uniform noise representation)
- Given PML simplicity, it is quite promising compared to STRAIGHT.

感謝您的關注

Gracias por su atención

Thank you for your attention

आप अपना ध्यान के लिए धन्यवाद

شكرا لكم على اهتمامكم

Obrigado pela sua atenção

Спасибо за ваше внимание

ご清聴ありがとう

మీ శ్రద్ధకు ధన్యవాదాలు

Je vous remercie de votre attention

Σας ευχαριστώ για την προσοχή σας

Dankon pro via atento