Vocoding approaches for statistical parametric speech synthesis

Ranniery Maia

Toshiba Research Europe Limited
Cambridge Research Laboratory

Speech Synthesis Seminar Series

CUED, University of Cambridge, UK

March 2nd, 2011

Topics of this presentation

1. Existing methods to generate the speech waveform in statistical parametric speech synthesis
2. An idea for closing the gap between acoustic modeling and waveform generation

Notation and acronyms in this presentation

Notation

$x(n)$	a discrete-time signal
$X(z)$	$x(n)$ in the z-transform domain
$X\left(e^{j\omega}\right)$	Discrete-Time Fourier Transform of $x(n)$ (frequency domain representation of $x(n)$)
$\left\|X\left(e^{j\omega}\right)\right\|$	magnitude response of $x(n)$
$\angle X\left(e^{j\omega}\right)$	phase response of $x(n)$
$\left\|X\left(e^{j\omega}\right)\right\|^2$	power spectrum of $x(n)$
$\boldsymbol{x}$	a vector
$\boldsymbol{X}$	a matrix

Acronyms

OLA	OverLap and Add
MELP	Mixed Excitation Linear Prediction
STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum
FFT	Fast Fourier Transform
IFFT	Inverse Fast Fourier Transform
LF	Liljencrants-Fant model
LP	Linear Prediction
PCA	Principal Component Analysis
LSP	Line Spectral Pairs

Contents

Contents

Introduction

Vocoding methods for statistical parametric speech synthesis
- Fully parametric excitation methods
- Methods that attempt to mimic the LP residual
- Methods that work on source and vocal tract modeling

Joint acoustic modeling and waveform generation for statistical parametric speech synthesis

Conclusion

Statistical parametric speech synthesis

- Speech synthesis methods
  1. Rule-based
     - 1.1 Parametric
     - 1.2 Unit concatenation
  2. Corpus-based
     - 2.1 Unit selection and concatenation
     - 2.2 Statistical parametric
     - 2.3 Hybrid

Statistical parametric speech synthesis

1. Advantages
    - several voices, small data, small footprint, language portability, etc
2. Unnatural synthesized speech
    2.1 Parametric model of speech production
    2.2 Parameters of the model are *averaged*

- How to alleviate this unnaturalness?
    1. Statistical modeling
    2. Choice of the speech production model
    3. Choice of the parameters to represent such model
    4. Way of synthesizing speech with these parameters

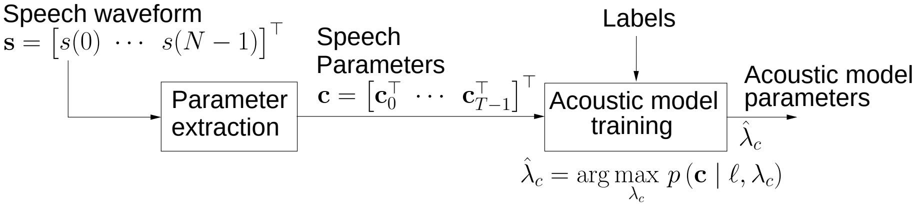
Statistical parametric speech synthesis

1. Advantages
   - several voices, small data, small footprint, language portability, etc
2. Unnatural synthesized speech
   2.1 Parametric model of speech production
   2.2 Parameters of the model are *averaged*

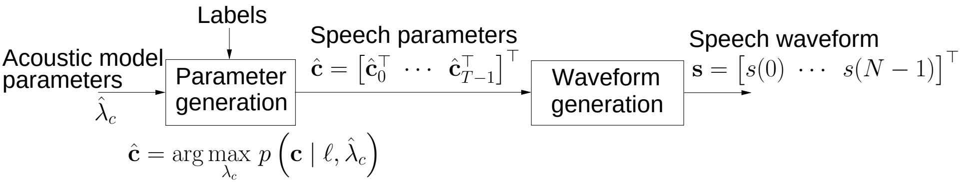
- How to alleviate this unnaturalness?
  1. Statistical modeling
  2. Choice of the speech production model
  3. Choice of the parameters to represent such model
  4. Way of synthesizing speech with these parameters

Statistical parametric speech synthesis

- Training time



- Synthesis time



Waveform generation part

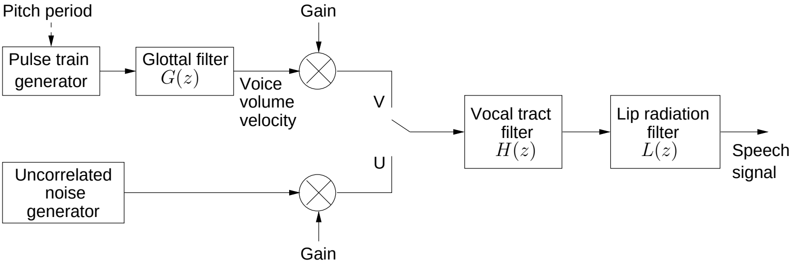
1. Choice of the speech production mechanism
   - Simple
     - Speech synthesis filter
     - Excitation
   - *Complete*
     - Vocal tract, glottal and lip radiation filters
     - Excitation
2. Appropriate parameters for the chosen speech mechanism
   - Good quantization/compression properties
3. Given the speech model and corresponding parameters, design the best way to synthesize the speech signal according to some criteria

Waveform generation part

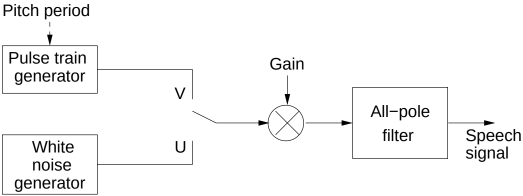
1. Choice of the speech production mechanism
   - Simple
     - Speech synthesis filter
     - Excitation
   - *Complete*
     - Vocal tract, glottal and lip radiation filters
     - Excitation
2. Appropriate parameters for the chosen speech mechanism
   - Good quantization/compression properties
3. Given the speech model and corresponding parameters, design the best way to synthesize the speech signal according to some criteria

Digital speech models

- The *complete model* [Deller, Jr. et al., 2000]

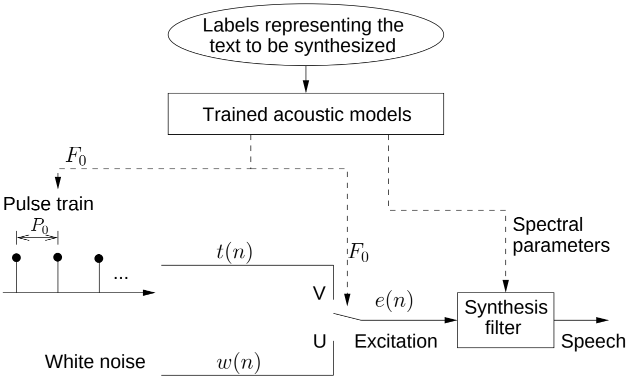


- The simplified model, assumed for LP analysis



Contents

Standard vocoder for statistical parametric synthesis



- Very simple
  - Analysis: F_0 extraction
  - Synthesis: pulse/white noise switch
- Poor speech quality!

Improved vocoding methods for statistical parametric synthesis

1. Methods that focus solely on the excitation signal
   1.1 Fully parametric excitation models
   1.2 Methods that attempt to mimic the LP residual
2. Methods that focus on source and vocal tract modeling

Contents

Introduction

Vocoding methods for statistical parametric speech synthesis

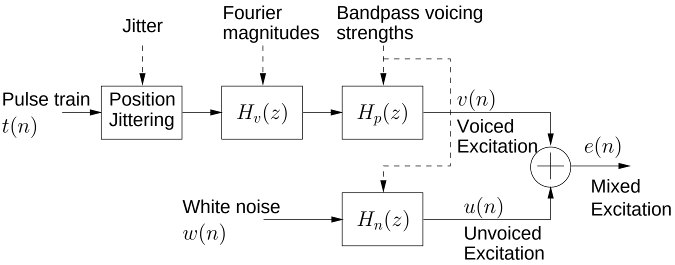
- Fully parametric excitation methods
- Methods that attempt to mimic the LP residual
- Methods that work on source and vocal tract modeling

Joint acoustic modeling and waveform generation for statistical parametric speech synthesis

Conclusion

MELP mixed excitation *[Yoshimura et al., 2001]*

MELP excitation building part



- Period jitter derived from voicing strengths for aperiodic frames
- Fourier magnitudes simulates the glottal filter
- Filters $H_p(z)$ and $H_n(z)$ control the amount of pulse and noise in the final excitation $e(n)$

Pulse and noise shaping filters

- Filters $H_p(z)$ and $H_n(z)$ *switch* between noise and pulse excitation according to each band

$$H_p(z) = \sum_{j=0}^{J-1} \sum_{m=0}^{M} \tilde{\beta}_j h_j(m) z^{-m} \, , \; H_n(z) = \sum_{j=0}^{J-1} \sum_{m=0}^{M} \left(1 - \tilde{\beta}_j\right) h_j(m) z^{-m}$$

$$\tilde{\beta}_j = \begin{cases} 1 & \text{if } \beta_j \geq 0.5 \\ 0 & \text{if } \beta_j < 0.5 \end{cases}$$

- $h_j(m)$: bandpass filter coefficients for the j band
- Bandpass voicing strength for the j band obtained according to a normalized correlation coefficient

$$\beta_j = f(r_t) \quad ; \quad r_t = \frac{\sum_{n=0}^{N-1} s(n)s(n+t)}{\sqrt{\left[\sum_{n=0}^{N-1} s^2(n)\right]\left[\sum_{n=0}^{N-1} s^2(n+t)\right]}}$$

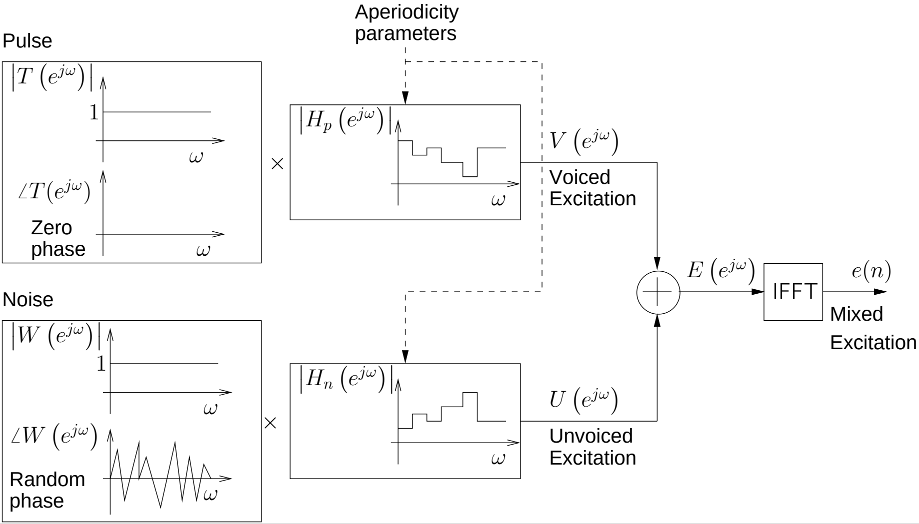
Application to statistical parametric synthesis

Additional parameters for acoustic modeling

1. Bandpass voicing strengths: 5
2. Fourier magnitudes: 10

STRAIGHT excitation *[Zen et al., 2007a]*

STRAIGHT vocoder: excitation construction ⇒ no phase manipulation case



STRAIGHT vocoder for statistical parametric synthesis

- Aperiodicity parameters extracted and averaged over specified frequency sub-bands
  - Band-aperiodicity parameters (BAP)
- At synthesis time the generated BAP are converted in aperiodicity
- Speech is synthesized in the frequency domain
- Achieves very good quality
- Additional parameters for acoustic modeling
  - BAP: usually 5 coefficients

Pulse and noise weighting filters

- Filters $H_p\left(e^{j\omega}\right)$ and $H_n\left(e^{j\omega}\right)$ shape the pulse and noise inputs, just like in MELP
- Frequency responses are obtained from the aperiodicity parameters $a(w)$

$$
\begin{aligned}
\left|H_p\left(e^{j\omega}\right)\right| &= \sqrt{1-a(w)} && 0 \leq \omega \leq \pi \\
\angle H_p\left(e^{j\omega}\right) &= 0 && 0 \leq \omega \leq \pi \\
\left|H_n\left(e^{j\omega}\right)\right| &= \sqrt{a(w)} && 0 \leq \omega \leq \pi \\
\angle H_n\left(e^{j\omega}\right) &= 0 && 0 \leq \omega \leq \pi
\end{aligned}
$$

Band aperiodicity parameters

- Aperiodicity at frequency ω

$$a(\omega) = \frac{\int w_{ERB}\left(\lambda;\omega\right)\left|S\left(e^{j\lambda}\right)\right|^2 \Upsilon\left(\frac{|S_L\left(e^{j\lambda}\right)|^2}{|S_U\left(e^{j\lambda}\right)|^2}\right) d\lambda}{\int w_{ERB}\left(\lambda;\omega\right)\left|S\left(e^{j\lambda}\right)\right|^2 d\lambda}$$

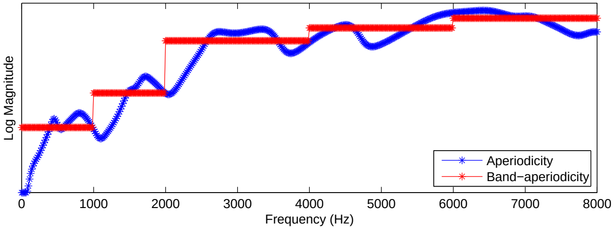
  - $\left|S\left(e^{j\omega}\right)\right|$: speech spectral envelope
  - $\left|S_U\left(e^{j\omega}\right)\right|$: envelope constructed by connecting the peaks of $\left|S\left(e^{j\omega}\right)\right|$
  - $\left|S_L\left(e^{j\omega}\right)\right|$: envelope constructed by connecting the valleys of $\left|S\left(e^{j\omega}\right)\right|$
  - $w_{ERB}\left(\lambda;\omega\right)$: auditory filter to smooth $\left|S\left(e^{j\omega}\right)\right|$
  - $\Upsilon(\cdot)$: look-up table operation
- Band-aperiodicity

$$b_j = \frac{1}{\Omega_j}\int_{\Omega_j} a(\omega)d\omega$$

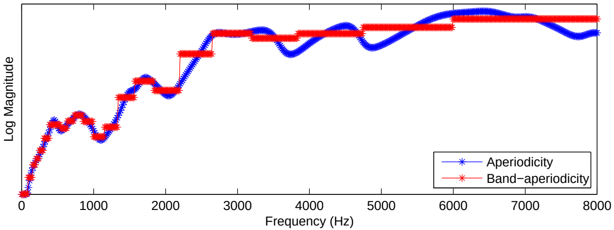
  - Ω_j: j-th frequency band

Aperiodicity and band aperiodicity: examples

- 5 bands: 0-1kHz, 1-2kHz, 2-4kHz, 4-6kHz, 6-8kHz

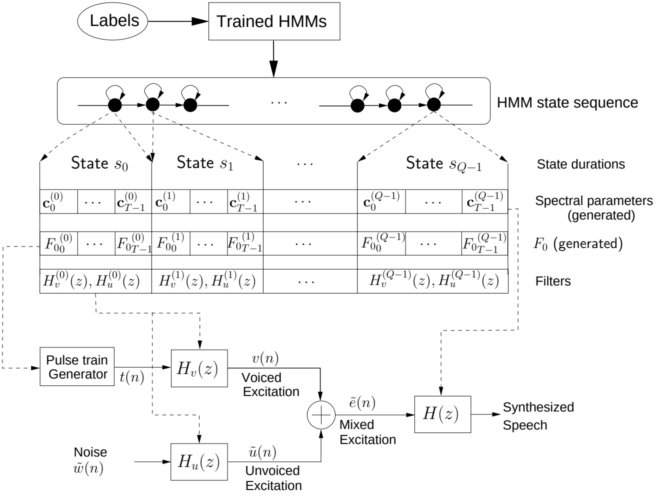


- 24 Bark critical bands



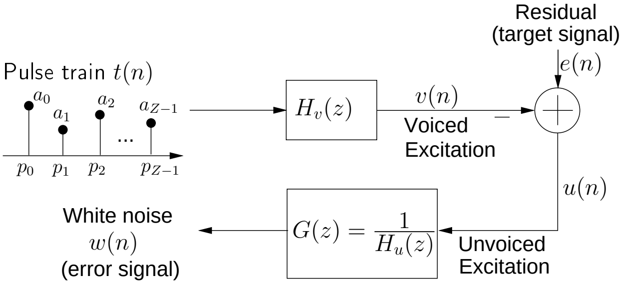
Contents

State-dependent mixed excitation *[Maia et al., 2007]*



Filters: $H_v(z) = \displaystyle\sum_{m=-M/2}^{M/2} h(m)z^{-m}$, $H_u(z) = \dfrac{1}{\sum_{l=0}^{L} g(l)z^{-l}}$

State-dependent mixed excitation: training



- Filter coefficients

$$\boldsymbol{h} = \left[h\left(-\frac{M}{2}\right) \quad \cdots \quad h\left(\frac{M}{2}\right)\right]^\top, \quad \boldsymbol{g} = \left[g(0) \quad \cdots \quad g(L)\right]^\top$$

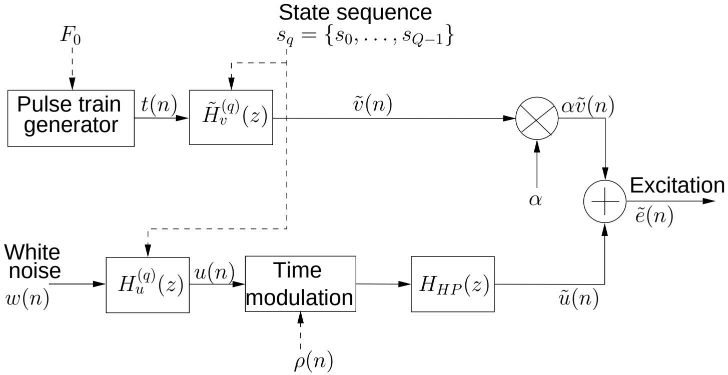
- And pulse positions and amplitudes

$$\{p_0, \ldots, p_{J-1}\}, \ \{a_0, \ldots, a_{J-1}\}$$

- Are optimized in a way that

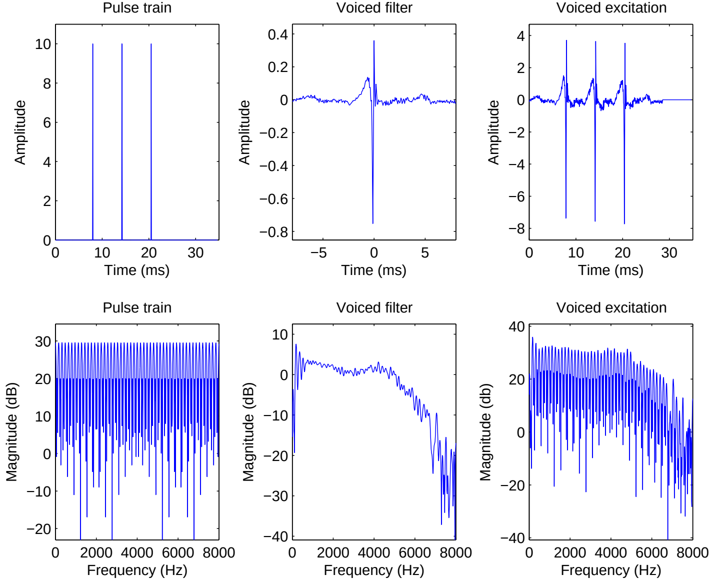
$$\{\boldsymbol{h}, \boldsymbol{g}, t(n)\} = \underset{\boldsymbol{g}, \boldsymbol{h}, t(n)}{\arg\max} \ P\left(e(n) \mid \boldsymbol{g}, \boldsymbol{h}, t(n)\right)$$

State-dependent mixed excitation: synthesis

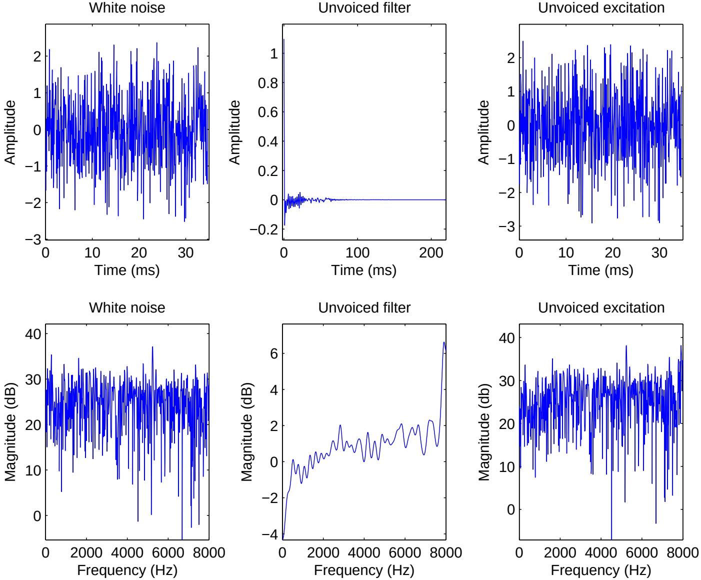


- Noise component is colored through
  1. High-pass filtering (F_c = 2kHz)
  2. Time modulation with a pitch-synchronous triangular window: $\rho(n)$
- $\tilde{H}_v(z)$ is normalized in energy
- Gain α adjusts the energy of the voiced component so that the power of the excitation signal $\tilde{e}(n)$ becomes one

Voiced filter effect



Unvoiced filter effect



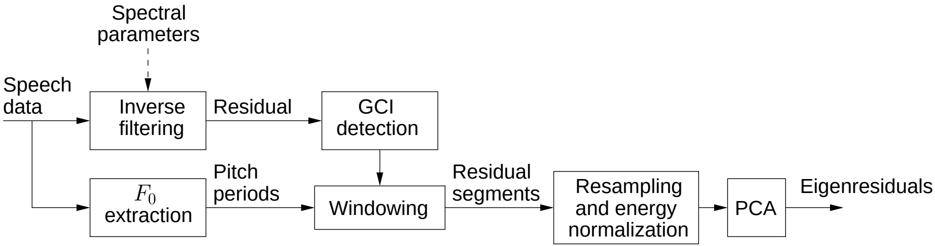
Deterministic plus stochastic residual modeling *[Drugman et al., 2009]*

- Assumed model of the LP residual $e(n)$

$$e(n) = e_d(n) + e_s(n)$$

  - $e_d(n)$: deterministic part
  - $e_s(n)$: stochastic part
- Maximum voiced frequency F_m
  - Boundary between deterministic and stochastic components
  - Set to 4 kHz

Deterministic modeling: *eigenresidual* calculation



- Normalized frequency F_0^* for resampling the residual segments

$$F_0^* \leq \frac{F_{\text{Nyquist}}}{F_m} F_{0,\text{min}}$$

- PCA: eigenresiduals explain about 80% of the total dispersion

TOSHIBA
Leading Innovation >>>

Stochastic modeling

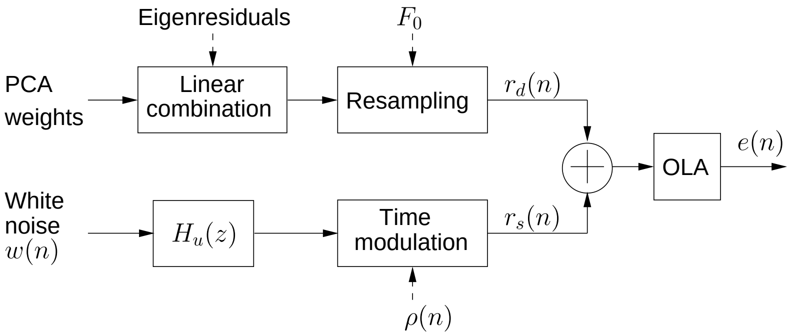
- Stochastic component model

$$e_s(n) = \rho(n)\,[h_u(n) * w(n)]$$

  - $\rho(n)$: pitch synchronous modulation window
  - $h_u(n)$: AR filter impulse response
  - $w(n)$: white noise
- Unvoiced filter $h_u(n)$
  - Fixed
  - Auto-regressive (all-pole)
  - Coefficients obtained through LP analysis

Application to statistical parametric synthesis

- Additional parameters for acoustic modeling
  - PCA weights: 15
- Use of eigenresidual of superior ranks makes no difference
  $\Longrightarrow$ Optionally, the stream of PCA weights can be removed
- Synthesis part



Waveform interpolation *[Sung et al., 2010]*

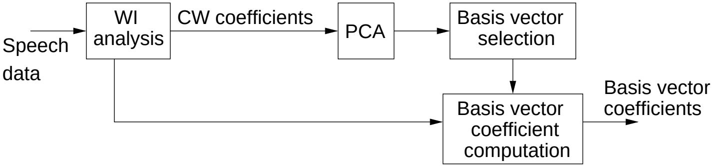
- Waveform interpolation (WI)
  - Each cycle of the excitation signal represented by a characteristic waveform (CW)

$$e(n) = \sum_{k=0}^{P/2} \left[A_k \cos\left(\frac{2\pi k n}{P}\right) + B_k \sin\left(\frac{2\pi k n}{P}\right) \right]$$

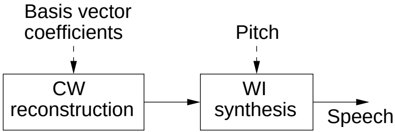
  - $\{A_k, B_k\}$: discrete-time Fourier series coefficients
  - P: pitch period
- CW extracted from the LP residual at a fixed rate
- Information to reconstruct the excitation signal
  - Pitch period: P
  - CW coefficients: $\{A_k, B_k\}$

Application to statistical parametric synthesis

- Analysis is similar to eigenresidual calculation [Drugman et al., 2009]



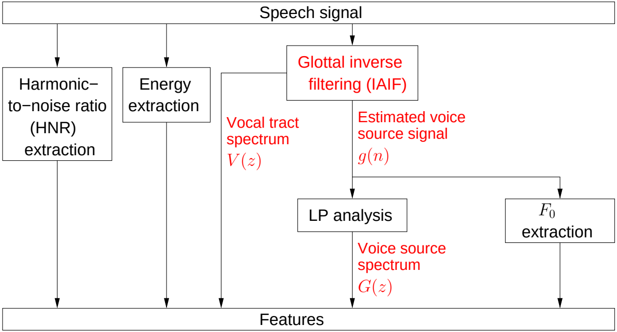
- Additional parameters for acoustic modeling
  - Coefficients of the basis vectors: 8
- Synthesis



Contents

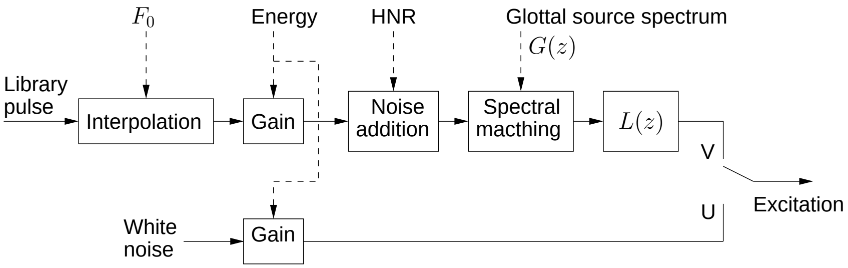
Glottal inverse filtering *[Raitio et al., 2008]*

- Uses Iterative Adaptive Inverse Filtering [Alku, 1992]



- Features for acoustic modeling
  1. F_0
  2. Energy
  3. HNR in 4 bands: 0-2kHz, 2-4kHz, 4-6kHz,6-8kHz
  4. Voice source spectrum $\Longrightarrow$ *glottal flow* $\Longrightarrow$ 10 LSPs
  5. Vocal tract spectrum: 30 LSPs

At synthesis time



- Library pulse extracted from the speech data through glottal inverse filtering
- Noise is separately added to each band of the voiced excitation according to HNR
- $G(z)$ implements the spectral shape of the glottal pulse
- $L(z)$ is a fixed lip radiation filter

Glottal spectrum separation (GSS) *[Cabral et al., 2008]*

- Speech production model

$$S\left(e^{j\omega}\right) = D\left(e^{j\omega}\right) G\left(e^{j\omega}\right) V\left(e^{j\omega}\right) R\left(e^{j\omega}\right)$$

$D\left(e^{j\omega}\right)$: pulse train	$G\left(e^{j\omega}\right)$: glottal pulse
$V\left(e^{j\omega}\right)$: vocal tract	$R\left(e^{j\omega}\right)$: lip radiation

- Simplified speech production model

$$S\left(e^{j\omega}\right) = D\left(e^{j\omega}\right) V\left(e^{j\omega}\right)$$

- What GSS does
  1. Estimate a model of the glottal flow derivative $\Longrightarrow E\left(e^{j\omega}\right)$
  2. Remove its effect from the speech spectral envelope $\Longrightarrow \hat{H}\left(e^{j\omega}\right)$

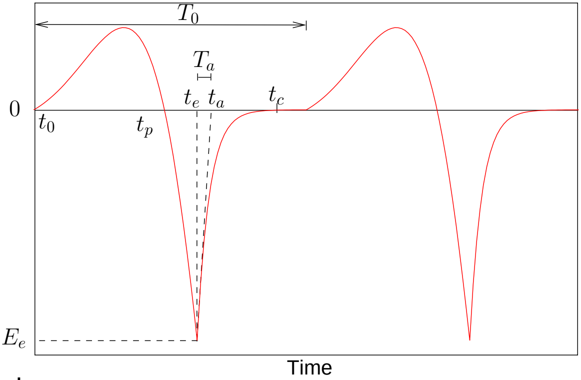
$$V\left(e^{j\omega}\right) = \frac{\hat{H}\left(e^{j\omega}\right)}{E\left(e^{j\omega}\right)}$$

  3. Re-synthesize speech

$$S\left(e^{j\omega}\right) = D\left(e^{j\omega}\right) E\left(e^{j\omega}\right) \frac{\hat{H}\left(e^{j\omega}\right)}{E\left(e^{j\omega}\right)} = S\left(e^{j\omega}\right) = D\left(e^{j\omega}\right) V\left(e^{j\omega}\right)$$

Glottal flow model utilized in GSS

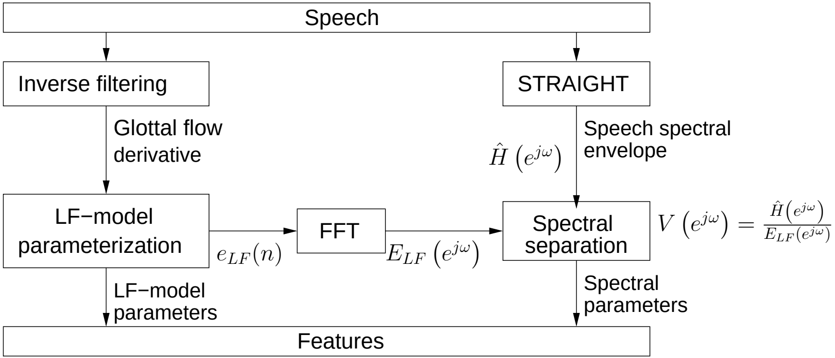
- LF model is used to represent the glottal pulse



- Parameters

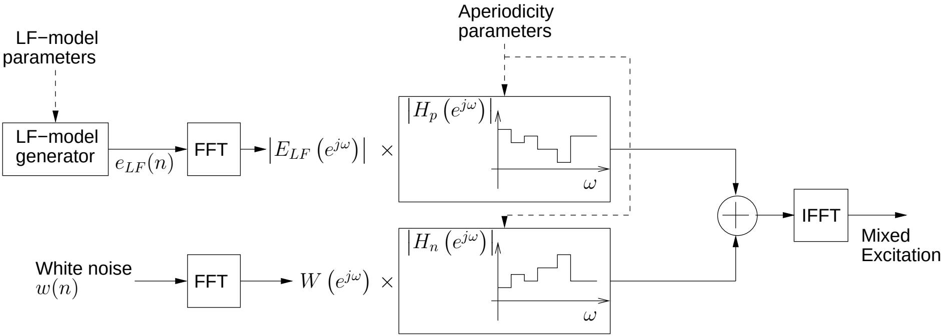
t_c: instant of complete closure	t_p: instant of maximum flow
t_e: instant of maximum excitation	$T_a = t_a - t_e$
T_0: fundamental period	E_e: amplitude of maximum excitation

Application to statistical parametric synthesis



- Acoustic modeling
  1. Spectral parameters: mel-cepstral coefficients
  2. Band aperiodicity parameters
  3. 5 LF-model parameters: t_e, t_p, T_a, E_e, T_0

Synthesis part



- STRAIGHT vocoder is utilized
- Original delta pulse is replaced by the pulse created by the generated LF-model parameters

Vocal tract and LF-model plus noise separation *[Lanchantin et al., 2010]*

- Assumed speech model

$$S\left(e^{j\omega}\right) = \left[D\left(e^{j\omega}\right) G\left(e^{j\omega}\right) + W\left(e^{j\omega}\right)\right] V\left(e^{j\omega}\right) R\left(e^{j\omega}\right)$$

$D\left(e^{j\omega}\right)$: pulse train	$G\left(e^{j\omega}\right)$: glottal pulse
$W\left(e^{j\omega}\right)$: white noise	$V\left(e^{j\omega}\right)$: vocal tract
$R\left(e^{j\omega}\right)$: lip radiation	

- Parameterization
  1. Fundamental frequency $\Rightarrow D\left(e^{j\omega}\right)$
  2. LF model parameter $\Rightarrow G\left(e^{j\omega}\right)$
  3. Noise power $\Rightarrow W\left(e^{j\omega}\right)$
  4. Mel-cepstral coefficients $\Rightarrow V\left(e^{j\omega}\right)$

Application to statistical parametric synthesis

1. Estimate a single parameter for a simplified LF model: R_d
2. Determine a maximum voiced frequency ω_{VU}
3. Estimate power of the noise $W\left(e^{j\omega}\right)$: σ_g^2
4. Estimate vocal tract parameters

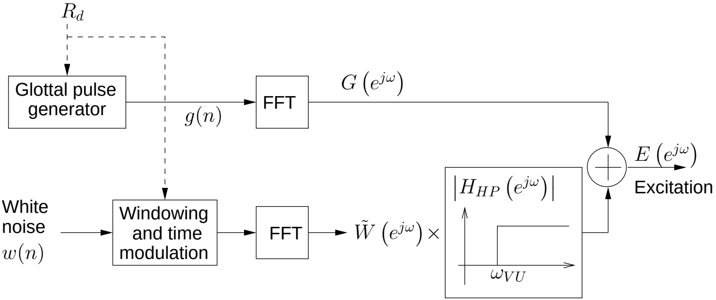
$$V\left(e^{j\omega}\right) = \begin{cases} \tau^o\left(\frac{S\left(e^{j\omega}\right)}{R\left(e^{j\omega}\right)G\left(e^{j\omega}\right)}\right)\gamma^{-1}, & w < \omega_{VU} \\ \mathcal{C}^o\left(\frac{S\left(e^{j\omega}\right)}{R\left(e^{j\omega}\right)G\left(e^{j\omega_{VU}}\right)}\right)\Gamma^{-1}, & w \geq \omega_{VU} \end{cases}$$

   - τ^o: cepstral analysis by fitting the harmonic peaks
   - $\mathcal{C}^o$: power cepstrum
   - Γ, γ: normalization terms

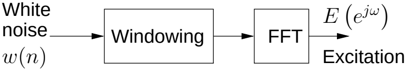
- Acoustic modeling
  1. Simplified LF-model parameter: R_d
  2. Standard deviation of the noise component: σ_g
  3. Cepstral coefficients that represent $V\left(e^{j\omega}\right)$
  4. F_0

Synthesis time

- Excitation construction for voiced segments



- Excitation construction for unvoiced segments



- Synthesized speech signal

$$S\left(e^{j\omega}\right) = E\left(e^{j\omega}\right) V\left(e^{j\omega}\right) R\left(e^{j\omega}\right)$$

Vocoding methods: summary and examples

Method	Description	Sample
Simple	Pulse train/white noise simple switch	
MELP	MELP mixed excitation	
STRAIGHT	STRAIGHT mixed excitation	
SDF	State-dependent filtering mixed excitation	
DSM	Deterministic plus stochastic of the residual	
WI	Waveform interpolation to statistical parametric synthesis	
GlottiHMM	Glottal inverse filtering	
GSS	Glottal source separation	
SVLN	Separation of vocal tract and LF-model plus noise	

Contents

Joint vocoding-acoustic modeling

- The goal of any speech synthesizer is to reproduce the *speech waveform*
  - Parameters of a *joint acoustic-excitation* model are estimated by maximizing the probability of the speech waveform

ℓ

Speech waveform $\mathbf{s} = \left[s(0) \;\; \cdots \;\; s(N-1) \right]^\top$ → Joint acoustic−excitation model training → $\hat{\lambda}$

$$\hat{\lambda} = \arg\max_{\lambda} p\left(\mathbf{s} \mid \ell, \lambda\right)$$

$$= \arg\max_{\lambda} \int p\left(\mathbf{s} \mid \mathbf{c}, \lambda\right) p\left(\mathbf{c} \mid \ell, \lambda\right) d\mathbf{c}$$

$\mathbf{c} = \left[\mathbf{c}_0^\top \;\; \cdots \;\; \mathbf{c}_{T-1}^\top \right]^\top$

Spectral parameter as hidden variable

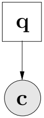
- $\lambda = \{\lambda_c, \lambda_e\}$: *acoustic-excitation* model parameters
  - λ_c: acoustic model part
  - λ_e: excitation model part

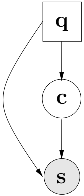
Another viewpoint: waveform-level modeling

- Comparison with typical modeling for parametric synthesis

Typical: state sequence is hidden variable, spectrum is the observation

New model: state sequence and spectrum are hidden variables, speech is the observation





- Similar concepts
  - [Toda and Tokuda, 2008]: factor analyzed trajectory HMM for spectral estimation
  - [Wu and Tokuda, 2009]: closed-loop training for HMM-based synthesis

A close look into the probabilities involved

- Typical statistical modeling for parametric synthesis

$$\hat{\lambda}_c = \mathop{\text{arg max}}_{\lambda_c} \sum_{\forall \boldsymbol{q}} p\left(\boldsymbol{c} \mid \boldsymbol{q}, \lambda_c\right) p\left(\boldsymbol{q} \mid \boldsymbol{\ell}, \lambda_c\right)$$

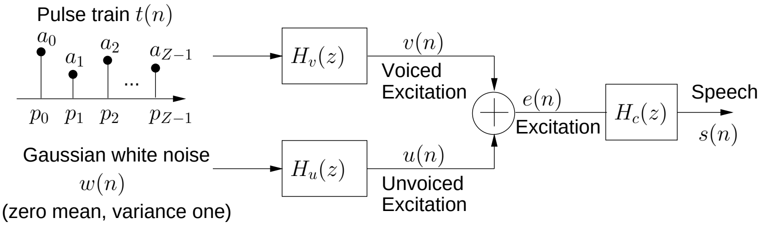
- *Augmented* statistical modeling

$$\hat{\lambda} = \mathop{\text{arg max}}_{\lambda} \sum_{\forall \boldsymbol{q}} \int \underbrace{p\left(\boldsymbol{s} \mid \boldsymbol{c}, \boldsymbol{q}, \lambda\right)}_{\Downarrow} \underbrace{p\left(\boldsymbol{c} \mid \boldsymbol{q}, \lambda\right) p\left(\boldsymbol{q} \mid \boldsymbol{\ell}, \lambda\right)}_{\Downarrow} d\boldsymbol{c}$$

Speech generative model (speech production from spectrum)

Can be modeled by existing machines, e.g. HMM, HSMM, trajectory HMM

One possible speech generative model



Probability of the speech signal

$$p\left(\boldsymbol{s} \mid \boldsymbol{H}_c, \boldsymbol{q}, \lambda_e\right) = |\boldsymbol{H}_c|^{-1} \mathcal{N}\left(\boldsymbol{H}_c^{-1}\boldsymbol{s}; \boldsymbol{H}_{v,\boldsymbol{q}}\boldsymbol{t}, \boldsymbol{\Phi}_{\boldsymbol{q}}\right)$$

$$\boldsymbol{\Phi}_{\boldsymbol{q}} = \left(\boldsymbol{G}_{\boldsymbol{q}}^{\top}\boldsymbol{G}_{\boldsymbol{q}}\right)^{-1}$$

$\lambda_e = \{\boldsymbol{H}_v, \boldsymbol{G}, \boldsymbol{t}\}$: excitation model parameters

Vocal tract filter impulse response and spectral parameters: relationship

- We need $p(\boldsymbol{s} \mid \boldsymbol{c}, \boldsymbol{q}, \lambda)$, not $p(\boldsymbol{s} \mid \boldsymbol{H}_c, \boldsymbol{q}, \lambda)$
  - **Mapping between $\boldsymbol{H}_c$ and $\boldsymbol{c}$ is necessary!**
- Two possibilities
  1. Relationship between $\boldsymbol{H}_c$ and $\boldsymbol{c}$ represented as a Gaussian process
  2. Relationship between $\boldsymbol{H}_c$ and $\boldsymbol{c}$ is deterministic
     - Cepstral coefficients!

Acoustic modeling

- **Trajectory HMM** [Zen et al., 2007b] for acoustic modeling

$$p(\boldsymbol{c} \mid \boldsymbol{\ell}, \lambda_c) = \sum_{\boldsymbol{q}} \underbrace{p(\boldsymbol{c} \mid \boldsymbol{q}, \lambda_c)}_{\downarrow \; \mathcal{N}(\boldsymbol{c}\,;\, \bar{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}})} \underbrace{p(\boldsymbol{q} \mid \boldsymbol{\ell}, \lambda_c)}_{\downarrow \; \pi_{q_0} \prod_{t=0}^{T-1} \alpha_{q_t q_{t+1}}}$$

$$\boxed{\begin{array}{c} \bar{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{r}_{\boldsymbol{q}} \\ \boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{P}_{\boldsymbol{q}}^{-1} = \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{W} \\ \boldsymbol{r}_{\boldsymbol{q}} = \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{\mu}_{\boldsymbol{q}} \\ \boldsymbol{W} : \text{append dynamic features to } \boldsymbol{c} \end{array}}$$

- **Why:** modeling of $p(\boldsymbol{c} \mid \boldsymbol{\ell}, \lambda_c)$ instead of $p(\boldsymbol{W}\boldsymbol{c} \mid \boldsymbol{\ell}, \lambda_c)$ (conventional HMM)

Joint acoustic-excitation model

Final joint model

$$p\left(\boldsymbol{s} \mid \boldsymbol{\ell}, \lambda\right) = \sum_{\boldsymbol{q}} \int \underbrace{p\left(\boldsymbol{s} \mid \boldsymbol{c}, \boldsymbol{q}, \lambda_e\right)} \underbrace{p\left(\boldsymbol{c} \mid \boldsymbol{q}, \lambda_c\right) p\left(\boldsymbol{q} \mid \boldsymbol{\ell}, \lambda_c\right)} d\boldsymbol{c}$$

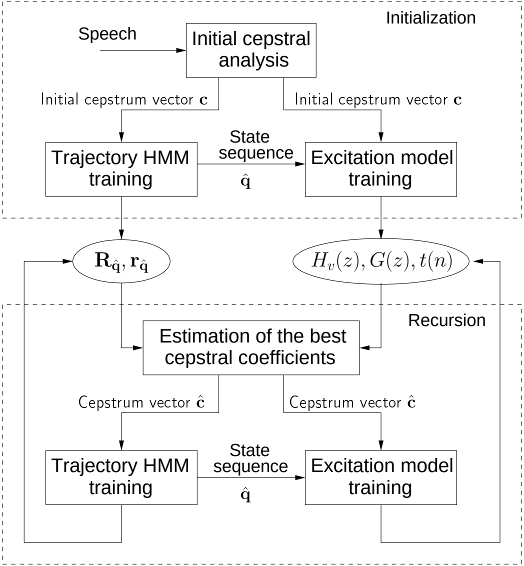
$\Downarrow$ Excitation model $\Downarrow$

$$|\boldsymbol{H}_c|^{-1} \mathcal{N}\left(\boldsymbol{H}_c^{-1}\boldsymbol{s}; \boldsymbol{H}_{v,\boldsymbol{q}}\boldsymbol{t}, \boldsymbol{\Phi}_{\boldsymbol{q}}\right)$$

$\Downarrow$ Acoustic model $\Downarrow$

$$\mathcal{N}\left(\boldsymbol{c}; \bar{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}}\right) \pi_{q_0} \prod_{t=0}^{T-1} \alpha_{q_t q_{t+1}}$$

Training procedure



Contents

Conclusions

- Quality improvement of statistical parametric synthesizers through better waveform generation methods
- Existing approaches that use source-filter modeling can be classified into
  - Methods that attempt to improve the excitation signal solely
  - Methods that focus on the speech production model as a whole
- Naturalness degradation of statistical parametric synthesizers has basically two causes
  - Acoustic modeling that produces averaged parameter trajectories
  - Use of parametric speech production models
- Methods which can integrate both acoustic modeling and speech production

Acknowledgments

Many thanks to

- **João Cabral,** from *University College Dublin, Ireland*
- **Tuomo Raitio,** from *Helsinki University of Technology, Finland*
- **Thomas Drugman,** from *Faculté Polytechnique de Mons, Belgium*
- **Pierre Lanchantin,** from *IRCAM, France*
- **June Sig Sung,** from *Seoul National University, South Korea*

for proving samples of their systems.

References I

Alku, P. (1992).
Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering.
Speech Communication, 11(2–3):109–118.

Cabral, J., Renals, S., Richmond, K., and Yamagishi, J. (2008).
Glottal spectral separation for parametric speech synthesis.
In *Proc. of Interspeech*, pages 1829–1832.

Deller, Jr., J. R., Hansen, J. H. L., and Proaks, J. G. (2000).
Discrete-Time Processing of Speech Signals.
IEEE Press Classic Reissue, New York.

Drugman, T., Wilfart, G., and Dutoit, T. (2009).
A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis.
In *Proc. of Interspeech*, pages 1779–1782.

Lanchantin, P., Degottex, G., and Rodet, X. (2010).
An HMM-based speech synthesis system using a new glottal source and vocal-tract separation method.
In *Proc. of ICASSP*, pages 4630–4633.

Maia, R., Toda, T., Zen, H., Nankaku, Y., and Tokuda, K. (2007).
An excitation model for HMM-based speech synthesis based on residual modeling.
In *Proc. of the 6th ISCA Workshop on Speesh Synthesis*, pages 131–136.

Raitio, T., Suni, A., Pulakka, H., Vainio, M., and Alku, P. (2008).
HMM-based Finnish text-to-speech system using glottal inverse filtering.
In *Interspeech*, pages 1881–1884.

References II

Sung, J., Kyung, D., Oh, H., and Kim, N. (2010).
Excitation modeling based on waveform interpolation for HMM-based speech synthesis.
In *Proc. of Interspeech*, pages 813–816.

Toda, T. and Tokuda, K. (2008).
Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM.
In *Proc. of ICASSP*, pages 3925–3928.

Wu, Y. J. and Tokuda, K. (2009).
Minimum generation error training by using original spectrum as reference for log spectral distortion measure.
In *Proc. of ICASSP*, pages 4013–4016.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2001).
Mixed-excitation for HMM-based speech synthesis.
In *Proc. of EUROSPEECH*.

Zen, H., Toda, T., Nakamura, M., and Tokuda, K. (2007a).
Details of the Nitech HMM-based speech synthesis for Blizzard Challenge 2005.
IEICE Trans. on Inf. and Systems, E90-D(1):325–333.

Zen, H., Tokuda, K., and Kitamura, T. (2007b).
Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequence.
Computer Speech and Language, 21(1):153–173.

TOSHIBA

Leading Innovation >>>