Speech Modifications

Lecture 3: Harmonic plus Noise Models and Modifications

Yannis Stylianou

Telecommunications and Network Lab, ICS- FORTH
styliano@ics.forth.gr

Vitoria, 2010 Sept 2nd

History

- MBE (Griffin et al.1988 [1])
- Sinusoids plus band-pass random signals (Abrantes et al.1991 [2]
- Harmonic and Stochastic Model (Laroche et al.1993 [3]
- Iterative decomposition of the excitation signal (Yegnayarayana et al.1995 [4])
- Harmonic plus Noise Model (Stylianou et al.1995 [5]
- Harmonic plus Noise Model 2 (Stylianou, PhD, 1996 [6])

History

- MBE (Griffin et al.1988 [1])
- Sinusoids plus band-pass random signals (Abrantes et al.1991 [2]
- Harmonic and Stochastic Model (Laroche et al.1993 [3]
- Iterative decomposition of the excitation signal (Yegnayarayana et al.1995 [4])
- Harmonic plus Noise Model (Stylianou et al.1995 [5]
- Harmonic plus Noise Model 2 (Stylianou, PhD, 1996 [6])

History

- MBE (Griffin et al.1988 [1])
- Sinusoids plus band-pass random signals (Abrantes et al.1991 [2]
- Harmonic and Stochastic Model (Laroche et al.1993 [3]
- Iterative decomposition of the excitation signal (Yegnayarayana et al.1995 [4])
- Harmonic plus Noise Model (Stylianou et al.1995 [5]
- Harmonic plus Noise Model 2 (Stylianou, PhD, 1996 [6])

History

- MBE (Griffin et al.1988 [1])
- Sinusoids plus band-pass random signals (Abrantes et al.1991 [2]
- Harmonic and Stochastic Model (Laroche et al.1993 [3]
- Iterative decomposition of the excitation signal (Yegnayarayana et al.1995 [4])
- Harmonic plus Noise Model (Stylianou et al.1995 [5]
- Harmonic plus Noise Model 2 (Stylianou, PhD, 1996 [6])

History

- MBE (Griffin et al.1988 [1])
- Sinusoids plus band-pass random signals (Abrantes et al.1991 [2]
- Harmonic and Stochastic Model (Laroche et al.1993 [3]
- Iterative decomposition of the excitation signal (Yegnayarayana et al.1995 [4])
- Harmonic plus Noise Model (Stylianou et al.1995 [5]
- Harmonic plus Noise Model 2 (Stylianou, PhD, 1996 [6])

History

- MBE (Griffin et al.1988 [1])
- Sinusoids plus band-pass random signals (Abrantes et al.1991 [2]
- Harmonic and Stochastic Model (Laroche et al.1993 [3]
- Iterative decomposition of the excitation signal (Yegnayarayana et al.1995 [4])
- Harmonic plus Noise Model (Stylianou et al.1995 [5]
- Harmonic plus Noise Model 2 (Stylianou, PhD, 1996 [6])

Why to decompose?

Decomposing speech into (quasi)periodic and non-periodic part has many applications in:

- Speech modification
- Speech coding
- Pathologic voice detection (i.e., HNR ...)
- Psychoacoustic research

Why to decompose?

Decomposing speech into (quasi)periodic and non-periodic part has many applications in:

- Speech modification
- Speech coding
- Pathologic voice detection (i.e., HNR ...)
- Psychoacoustic research

Why to decompose?

Decomposing speech into (quasi)periodic and non-periodic part has many applications in:

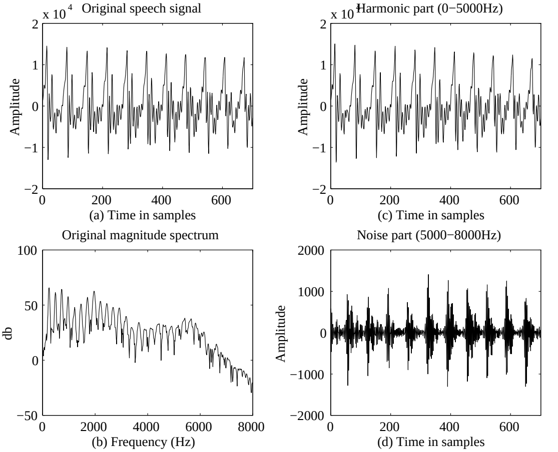
- Speech modification
- Speech coding
- Pathologic voice detection (i.e., HNR ...)
- Psychoacoustic research

Why to decompose?

Decomposing speech into (quasi)periodic and non-periodic part has many applications in:

- Speech modification
- Speech coding
- Pathologic voice detection (i.e., HNR ...)
- Psychoacoustic research

Motivation for HNM



Brief overview of HNM

- HNM is a pitch-synchronous harmonic plus noise representation of the speech signal.
- Speech spectrum is divided into a low and a high band delimited by the so-called maximum voiced frequency.
- The low band of the spectrum (below the maximum voiced frequency) is represented solely by harmonically related sine waves.
- The upper band is modeled as a noise component modulated by a time-domain amplitude envelope.
- HNM allows high-quality copy synthesis and prosodic modifications.

HNM IN EQUATIONS

- Harmonic part:

$$h(t) = \sum_{k=-L(t)}^{L(t)} A_k(t) e^{j\,k\omega_0(t)\,t}$$

- Noise part:

$$n(t) = e(t)\,[v(\tau, t) \star b(t)]$$

- Speech:

$$s(t) = h(t) + n(t)$$

Periodic part

- HNM_1: Sum of exponential functions without slope

$$h_1[n] = \sum_{k=-L(n_a^i)}^{L(n_a^i)} a_k(n_a^i) e^{j2\pi k f_0(n_a^i)(n-n_a^i)}$$

- HNM_2: Sum of exponential function with complex slope

$$h_2[n] = \Re\left\{\sum_{k=1}^{L(n_a^i)} A_k(n) \exp^{j2\pi k f_0(n_a^i)(n-n_a^i)}\right\}$$

where

$$A_k(n) = a_k(n_a^i) + (n - n_a^i) b_k(n_a^i)$$

with $a_k(n_a^i), b_k(n_a^i)$ to be complex numbers (amplitude and slope respectively). $\Re$ denotes taking the real part.

Periodic part

- HNM_1: Sum of exponential functions without slope

$$h_1[n] = \sum_{k=-L(n_a^i)}^{L(n_a^i)} a_k(n_a^i) e^{j2\pi k f_0(n_a^i)(n-n_a^i)}$$

- HNM_2: Sum of exponential function with complex slope

$$h_2[n] = \Re\left\{ \sum_{k=1}^{L(n_a^i)} A_k(n) \exp^{j2\pi k f_0(n_a^i)(n-n_a^i)} \right\}$$

where

$$A_k(n) = a_k(n_a^i) + (n - n_a^i) b_k(n_a^i)$$

with $a_k(n_a^i), b_k(n_a^i)$ to be complex numbers (amplitude and slope respectively). $\Re$ denotes taking the real part.

Periodic part *continuing*

- HNM_3: Sum of sinusoids with time-varying real amplitudes

$$h_3[n] = \sum_{k=0}^{L(n_a^i)} a_k(n) cos(\varphi_k(n))$$

where

$$\begin{aligned} a_k(n) &= c_{k0} + c_{k1}\,(n - n_a^i)^1 + \cdots + c_{kp}\,(n - n_a^i)^{p(n)} \\ \varphi_k(n) &= \epsilon_k + 2\pi\, k\zeta\,(n - n_a^i) \end{aligned}$$

where $p(n)$ is the order of the amplitude polynomial, which is, in general, a time-varying parameter.

Residual (Noise) part

The non-periodic part is just the *residual* signal obtained by subtracting the periodic-part (harmonic part) from the original speech signal in the time-domain

$$r[n] = s[n] - h[n]$$

where $h[n]$ is either $h_1[n], h_2[n]$, or $h_3[n]$.

Parameters Estimation

- Pitch: based on an autocorrelation criterion
- Maximum Voiced Frequency: voicing matching criteria
- Amplitudes: Least-Squares
- Noise part: variance and Linear prediction

Parameters Estimation

- Pitch: based on an autocorrelation criterion
- Maximum Voiced Frequency: voicing matching criteria
- Amplitudes: Least-Squares
- Noise part: variance and Linear prediction

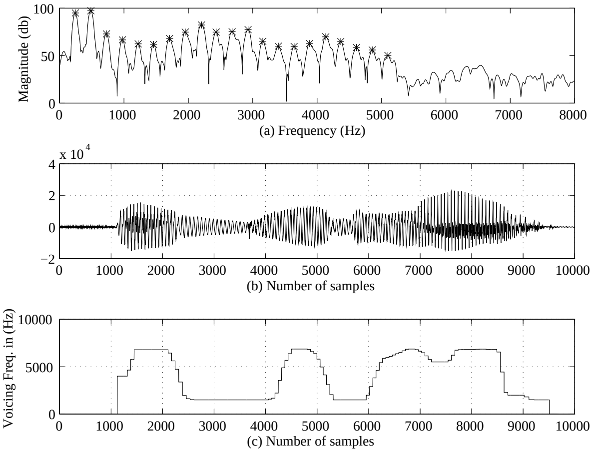
Parameters Estimation

- Pitch: based on an autocorrelation criterion
- Maximum Voiced Frequency: voicing matching criteria
- Amplitudes: Least-Squares
- Noise part: variance and Linear prediction

Parameters Estimation

- Pitch: based on an autocorrelation criterion
- Maximum Voiced Frequency: voicing matching criteria
- Amplitudes: Least-Squares
- Noise part: variance and Linear prediction

Example of Estimation

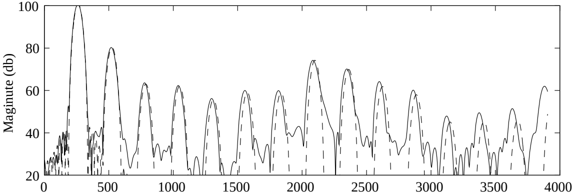


Fundamental frequency refinement

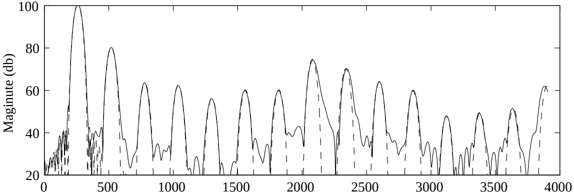
Using the initial f_0 value and the L detected voiced frequencies f_i, then the refined fundamental frequency, $\hat{f}_0$ is defined as the value that minimizes the error:

$$E(\hat{f}_0) = \sum_{i=1}^{L} |f_i - i \cdot \hat{f}_0|^2$$

Refinement frequency example

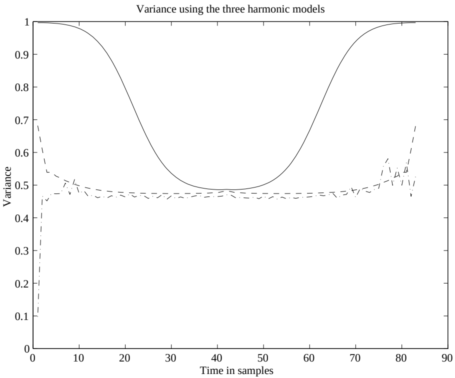


(a) Frequency in Hz



(b) Frequency in Hz

Variance of the residual signal



Synthesis

- Direct time frequency matching
- For the periodic part: Overlap-and-Add, or
  - Linear amplitude interpolation
  - Linear phase interpolation using average pitch value
- For the stochastic (noise) part):
  - Instead of AR coefficients we use reflection coefficients
  - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
  - Modulation in time with a deterministic function (i.e., triangular)

Synthesis

- Direct time frequency matching
- For the periodic part: Overlap-and-Add, or
  - Linear amplitude interpolation
  - Linear phase interpolation using average pitch value
- For the stochastic (noise) part):
  - Instead of AR coefficients we use reflection coefficients
  - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
  - Modulation in time with a deterministic function (i.e., triangular)

Synthesis

- Direct time frequency matching
- For the periodic part: Overlap-and-Add, or
  - Linear amplitude interpolation
  - Linear phase interpolation using average pitch value
- For the stochastic (noise) part):
  - Instead of AR coefficients we use reflection coefficients
  - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
  - Modulation in time with a deterministic function (i.e., triangular)

Synthesis

- Direct time frequency matching
- For the periodic part: Overlap-and-Add, or
  - Linear amplitude interpolation
  - Linear phase interpolation using average pitch value
- For the stochastic (noise) part):
  - Instead of AR coefficients we use reflection coefficients
  - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
  - Modulation in time with a deterministic function (i.e., triangular)

Synthesis

- Direct time frequency matching
- For the periodic part: Overlap-and-Add, or
  - Linear amplitude interpolation
  - Linear phase interpolation using average pitch value
- For the stochastic (noise) part):
  - Instead of AR coefficients we use reflection coefficients
  - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
  - Modulation in time with a deterministic function (i.e., triangular)

Synthesis

- Direct time frequency matching
- For the periodic part: Overlap-and-Add, or
  - Linear amplitude interpolation
  - Linear phase interpolation using average pitch value
- For the stochastic (noise) part):
  - Instead of AR coefficients we use reflection coefficients
  - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
  - Modulation in time with a deterministic function (i.e., triangular)

Synthesis

- Direct time frequency matching
- For the periodic part: Overlap-and-Add, or
  - Linear amplitude interpolation
  - Linear phase interpolation using average pitch value
- For the stochastic (noise) part):
  - Instead of AR coefficients we use reflection coefficients
  - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
  - Modulation in time with a deterministic function (i.e., triangular)

Synthesis

- Direct time frequency matching
- For the periodic part: Overlap-and-Add, or
  - Linear amplitude interpolation
  - Linear phase interpolation using average pitch value
- For the stochastic (noise) part):
  - Instead of AR coefficients we use reflection coefficients
  - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
  - Modulation in time with a deterministic function (i.e., triangular)

Preparing for prosody modifications

For being able to make prosodic modifications (at least pitch and time scale modifications) in the context of HNM, we should

- Associate analysis and synthesis time instants
- Determine a continuous magnitude envelope
- Determine a continuous phase envelope

Preparing for prosody modifications

For being able to make prosodic modifications (at least pitch and time scale modifications) in the context of HNM, we should

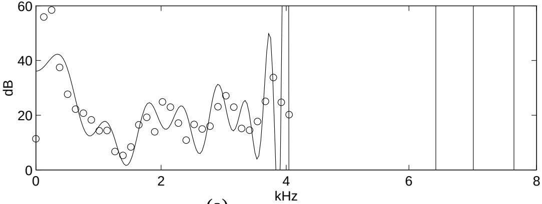
- Associate analysis and synthesis time instants
- Determine a continuous magnitude envelope
- Determine a continuous phase envelope

Preparing for prosody modifications

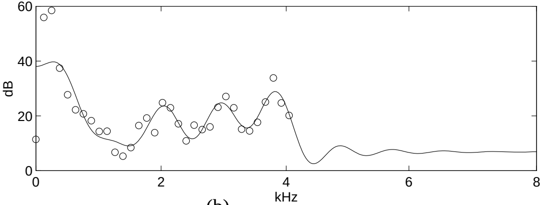
For being able to make prosodic modifications (at least pitch and time scale modifications) in the context of HNM, we should

- Associate analysis and synthesis time instants
- Determine a continuous magnitude envelope
- Determine a continuous phase envelope

Magnitude: Discrete Cepstrum

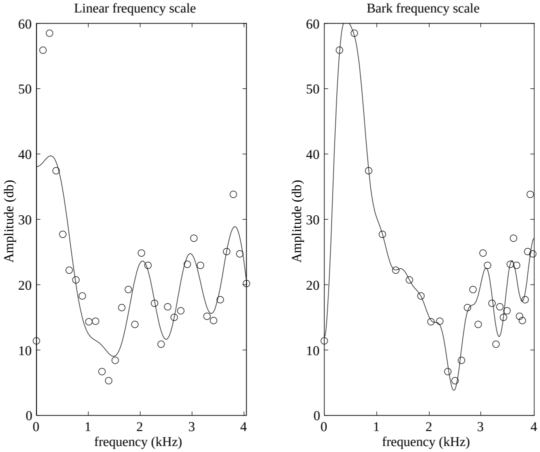


(a)

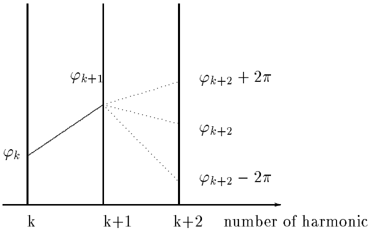


(b)

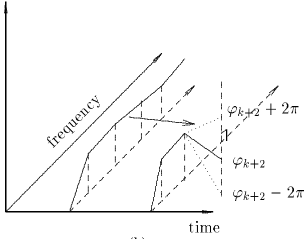
Bark scaled Disc. Cepstra



Phase envelope estimation

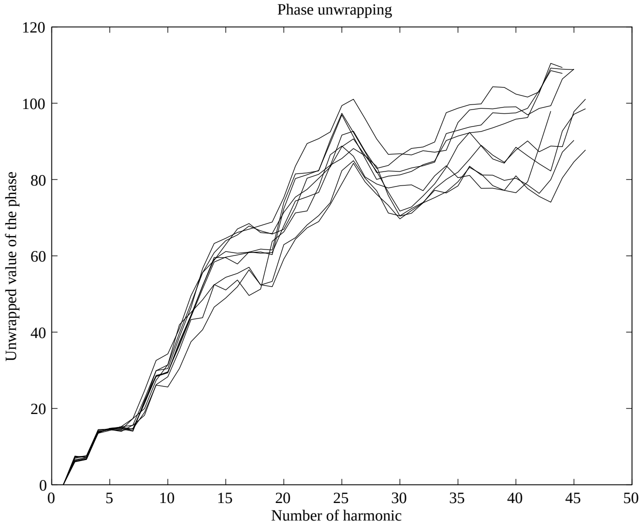


(a)

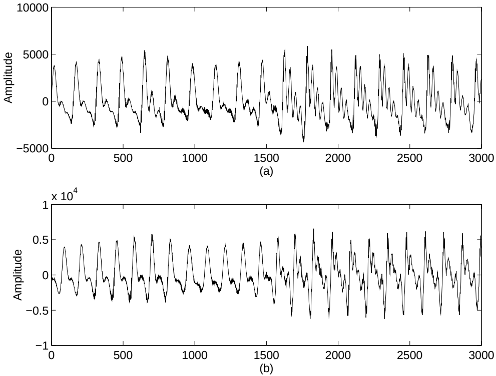


(b)

Example of Phase envelopes



Example of pitch modification



Pitch modification by 1.3

Sound Examples

Original	
Time-scale by 0.7	
Time-scale by 1.6	
Pitch modification by 0.8	
Pitch modification by 1.6	
Original	
Time-varying pitch and time modif.	
Original	
Time-scale by 4	
Time-scale by 6	

Probabilistic classification

- Modeling of the acoustic space of a speaker by a GMM :

$$p(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),$$

- Classification:

$$P(\mathcal{C}_i|\mathbf{x}) = \frac{\alpha_i N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{j=1}^{m} \alpha_j N(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

- Estimation using an Expectation-Maximization (EM) algorithm initialized by a standard binary splitting VQ procedure.

Mapping function

- Mapping function [7]:

$$\mathcal{F}(\mathbf{x}_t) = \sum_{i=1}^{m} P(\mathcal{C}_i|\mathbf{x}_t)\left[\boldsymbol{\nu}_i + \boldsymbol{\Gamma}_i \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_i)\right]$$

- Motivation:

$$E[\mathbf{y}|\mathbf{x} = \mathbf{x}_t] = \nu + \boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_t - \mu)$$

- Estimation of mapping function:

$$\epsilon = \sum_{t=1}^{n} ||\mathbf{y}_t - \mathcal{F}(\mathbf{x}_t)||^2$$

Mapping function

- Mapping function [7]:

$$\mathcal{F}(\mathbf{x}_t) = \sum_{i=1}^{m} P(\mathcal{C}_i|\mathbf{x}_t)\left[\boldsymbol{\nu}_i + \boldsymbol{\Gamma}_i\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_i)\right]$$

- Motivation:

$$E[\mathbf{y}|\mathbf{x} = \mathbf{x}_t] = \boldsymbol{\nu} + \boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_t - \boldsymbol{\mu})$$

- Estimation of mapping function:

$$\epsilon = \sum_{t=1}^{n} ||\mathbf{y}_t - \mathcal{F}(\mathbf{x}_t)||^2$$

Mapping function

- Mapping function [7]:

$$\mathcal{F}(\mathbf{x}_t) = \sum_{i=1}^{m} P(\mathcal{C}_i|\mathbf{x}_t)\left[\boldsymbol{\nu}_i + \boldsymbol{\Gamma}_i\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_i)\right]$$

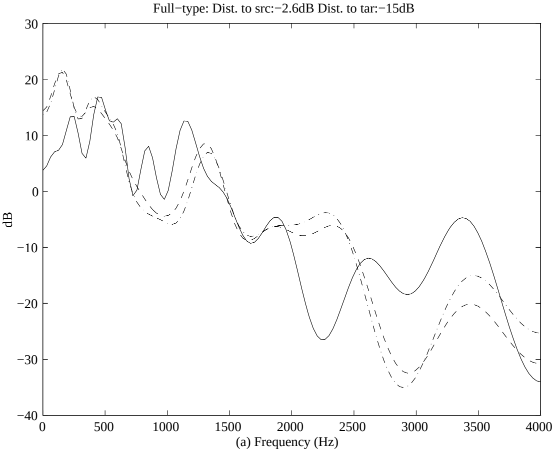
- Motivation:

$$E[\mathbf{y}|\mathbf{x} = \mathbf{x}_t] = \boldsymbol{\nu} + \boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_t - \boldsymbol{\mu})$$

- Estimation of mapping function:

$$\epsilon = \sum_{t=1}^{n} ||\mathbf{y}_t - \mathcal{F}(\mathbf{x}_t)||^2$$

Performance of the mapping function



Results - XAB test

Task: Listeners were asked to select either A or B as being most similar to X.

	PO	16 GMM	64 GMM	64 GMM(2)
Correct answers	18%	83%	88%	97%

Audio examples of Voice Conversion: HNM + GMM

Source	Converted	Target
Source	Converted	Target

D. Griffin and J. Lim, "Multiband-excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 236–243, Fev 1988.

A. Abrantes, J. Marques, and I. Transcoso, "Hybrid sinusoidal modeling of speech without voicing decision," *Eurospeech-91*, pp. 231–234, 1991.

J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model.," *Proc. IEEE ICASSP-93, Minneapolis*, pp. 550–553, Apr 1993.

B.Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," vol. 6, no. 1, 1998.

Y. Stylianou, J. Laroche, and E. Moulines, "High-Quality Speech Modification based on a Harmonic + Noise Model.," *Proc. EUROSPEECH*, 1995.

Y. Stylianou, *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification.*
PhD thesis, Ecole Nationale Supèrieure des Télécommunications, Jan 1996.

Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.