ICSP 2008

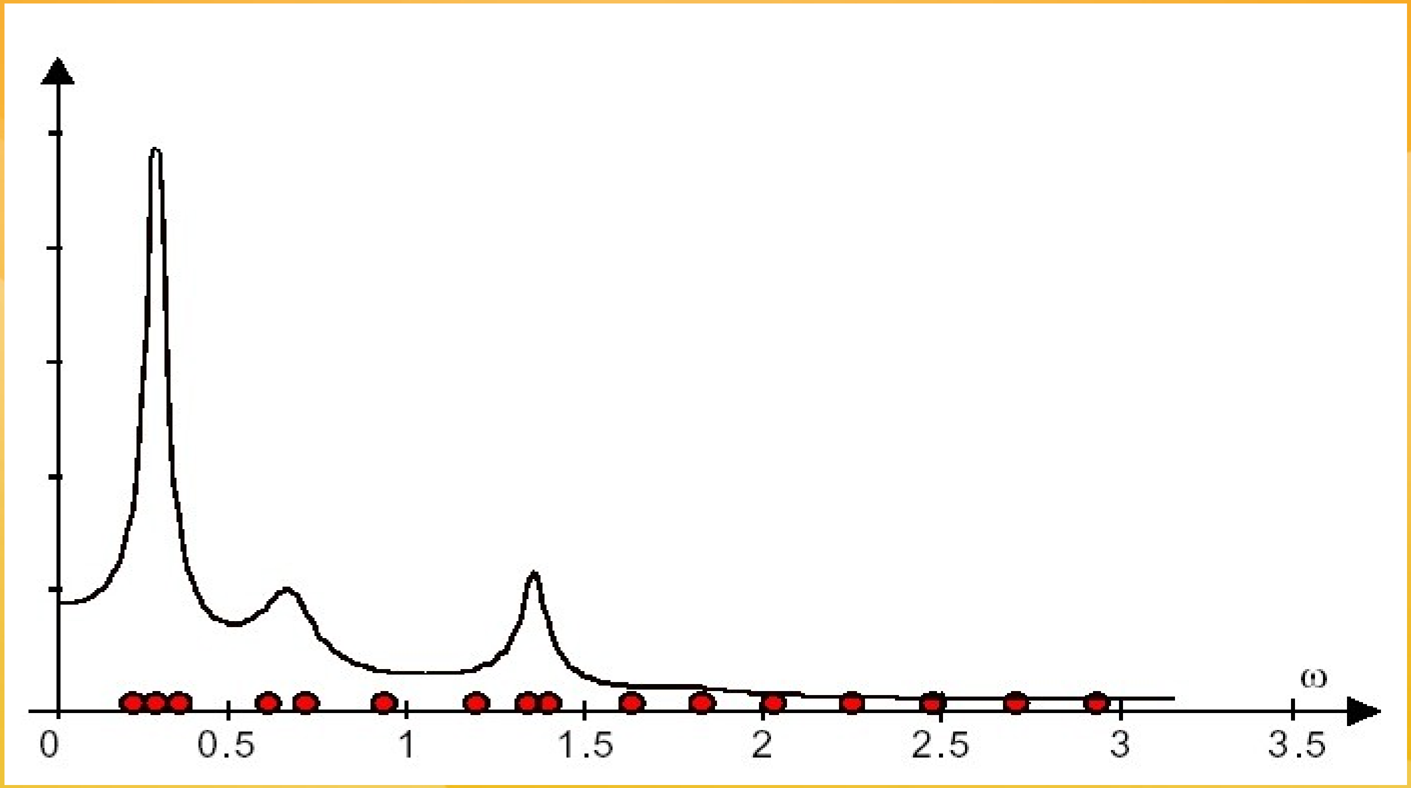
On Using Warping Function for LSFs Transformation in a Voice Conversion System

Zdeněk Hanzlíček and Jindřich Matoušek
University of West Bohemia, Czech Republic

Line Spectral Frequencies

- LSF = alternative representation of LPC
- position of particular line spectral frequency (LSF)

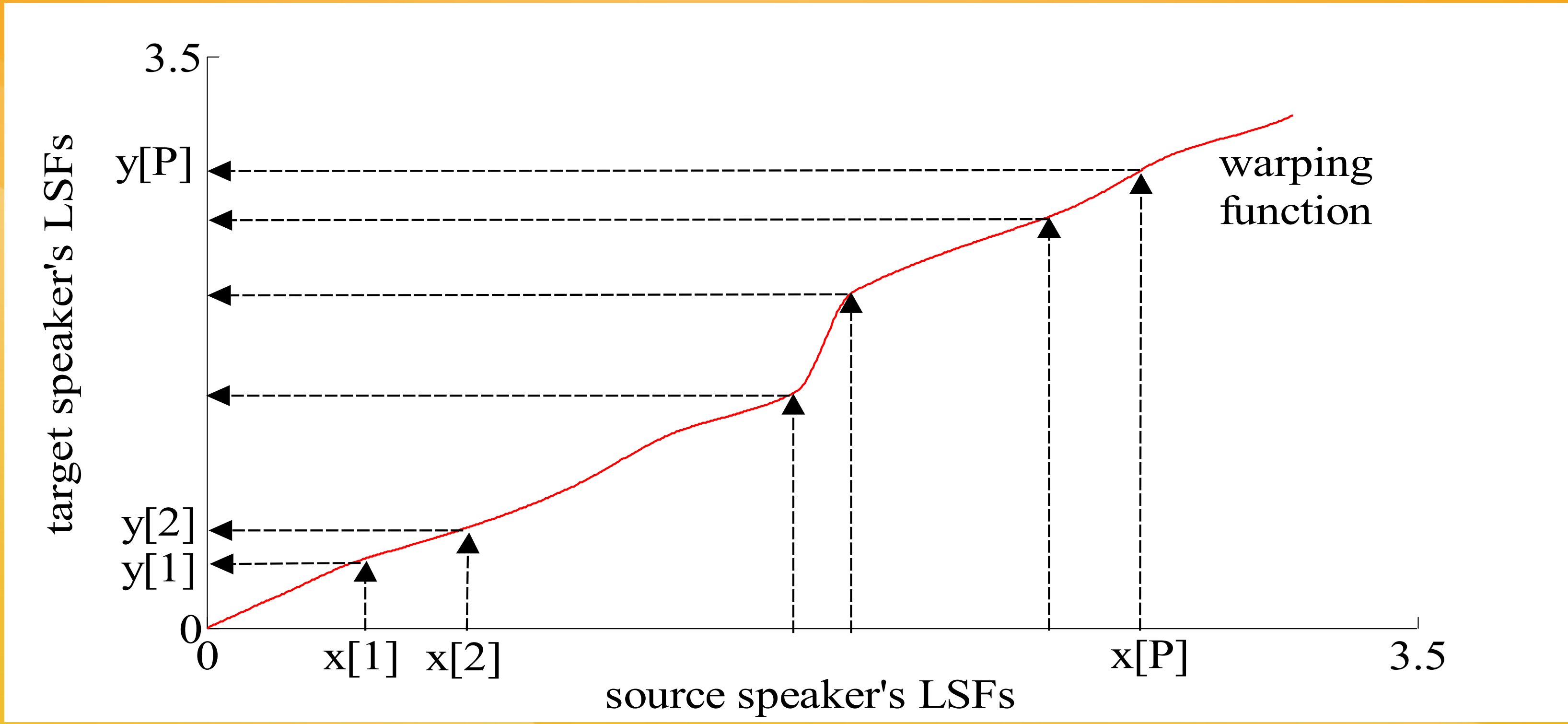
  => shape of spectral envelope



Main idea

LSFs shifting ~ spectral envelope transformation

=> shifting by frequency axis warping



LSF warping function (1)

- clustering of joint LSF vector $z = [x^T, y^T]^T$ into K classes (bisective k-means)
  - mean vector

$$\mu_z^k = \left[\mu_x^k[1], \ldots \mu_x^k[P], \mu_y^k[1], \ldots \mu_y^k[P]\right]^T$$

  - covariance matrix

$$\Sigma_z^k = diag\left[\sigma_x^k[1], \ldots \sigma_x^k[P], \sigma_y^k[1], \ldots \sigma_y^k[P]\right]$$

  - warping function

$$\tilde{y}[i] = f_k(x[i])$$

- resulting conversion function

$$\tilde{y}[i] = \sum_{k=1}^{K} w_k(x) f_k(x[i])$$

LSF warping function (2)

- requirements for warping function of k-th class

$$f_k(\mu_x^k[j]) = \mu_y^k[j]$$

- in each interval

$$\left\langle \mu_x^k[j], \mu_x^k[j+1] \right\rangle$$

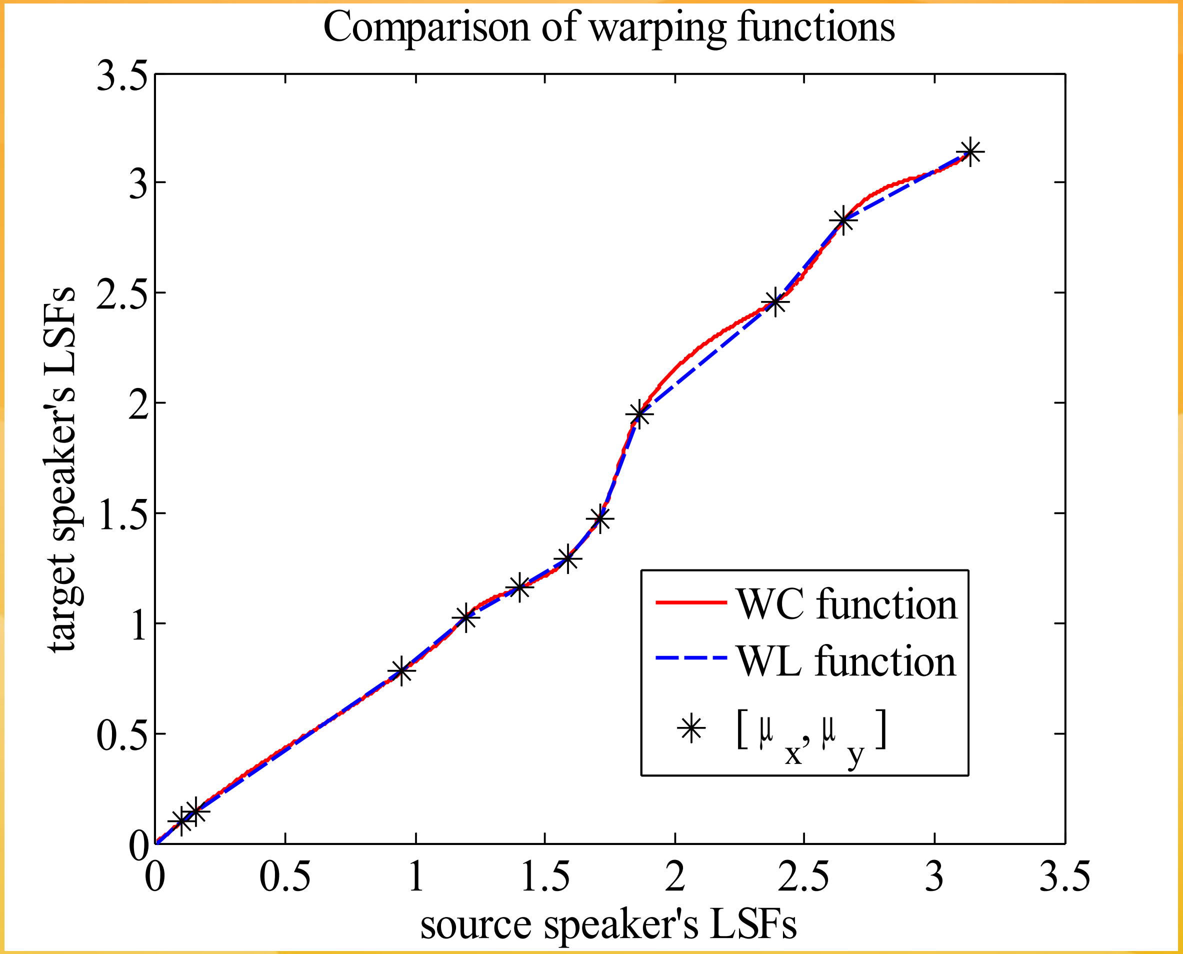
warping function is partly linear (WL function)

$$\tilde{y}[i] = a_k^j x[i] + b_k^j$$

or partly cubic (WC function)

$$\tilde{y}[i] = a_k^j x^3[i] + b_k^j x^2[i] + c_k^j x[i] + d_k^j$$

LSF warping function (3)



Experiments

- voice conversion system
  - true envelope estimator => LSFs
    - GMM-based transformation
    - warping function
  - pitch transformed by Gaussian normalisation
  - spectral detail transformed by residual prediction
- training set – 40 short sentences
- testing set – 10 sentences

Objective evaluation (1)

- performance index $$P_{LSF}=1-\frac{\sum_{n=1}^{N} d(\tilde{y}_n, y_n)}{\sum_{n=1}^{N} d(x_n, y_n)}$$
- higher value => better similarity

Function	Number of classes	Target speaker			
		M1	M2	F1	F2
GMM	5	0.424	0.326	0.303	0.346
	10	0.425	0.328	0.301	0.345
WC	5	0.235	0.149	0.135	0.184
	10	0.264	0.172	0.153	0.203
	150	0.314	0.219	0.189	0.240

Objective evaluation (2)

- global variance ratio $$R_{GV} = \frac{1}{P} \sum_{p=1}^{P} \frac{GV(\tilde{y}[p])}{GV(y[p])}$$

- value closer to 1 => better quality

Function	Number of classes	Target speaker			
		M1	M2	F1	F2
GMM	5	0.642	0.676	0.670	0.686
	10	0.655	0.691	0.700	0.683
WC	5	1.081	0.968	1.004	0.958
	10	1.038	0.953	1.004	0.958
	150	0.935	0.912	0.955	0.912

Subjective evaluation

- listening test – 10 participants
- 20 quintuples of utterances:
  - source, target, 3 transformed
- average order (lower value => better similarity / quality)
- GMM => better similarity
- WC => better quality

	GMM	combined	WC
Similarity	1.83 ± 0.56	1.94 ± 0.31	2.23 ± 0.59
Quality	2.34 ± 0.55	2.02 ± 0.45	1.65 ± 0.62

Thank you for your attention.