TRANSMUTATIVE VOICE CONVERSION

The quest for improving voice conversion quality

S.H. Mohammadi and A. Kain

Center for Spoken Language Understanding (CSLU)
Oregon Health & Science University (OHSU)
Portland, Oregon, USA

May 27, 2013

Introduction

- Voice Conversion: making the speech of a source speaker to sound like a target speaker
- Important Concerns:
  - Speaker Recognizability
  - Speaker Quality

Introduction

- Two main approaches:
  - Generative: typically uses a compact parametrization of speech and maps input to output parameters directly
  - Transmutative: modifies high-dimensional features of a high-fidelity speech model, leaving critical details unmodified
- Original, LPC Vocoded and Harmonic Vocoded speech

Introduction

- Unparallel Corpus: Given parallel (same-content) feature sequence $\tilde{X}_{N_s \times d}$ and $\tilde{Y}_{N_t \times d}$
- Align them: using Dynamic Time Warping (DTW)
- Parallelized Corpus: N frames of time-aligned features $X^{train}_{N \times d}$ and $Y^{train}_{N \times d}$

Generative Approach

- During Training: we find the optimal parameter set

$$\lambda^* = \arg\min_{\lambda} \mathrm{E}\left(\mathrm{Y}^{\mathrm{train}}, \mathcal{F}(\mathrm{X}^{\mathrm{train}}, \lambda)\right) \tag{1}$$

- During Conversion: features are mapped by evaluating

$$\mathcal{F}(\mathrm{X}^{\mathrm{test}}, \lambda^*) = \hat{\mathrm{Y}}^{\mathrm{test}}, \tag{2}$$

- Features: Line Spectral Frequency (LSF) in this study
- Mapping Function: Joint-Density Gaussian Mixture Model

Generative Approach

- These methods usually do not have a high quality because they are transformed to a lower dimension domain.
- Re-synthesis from them results in low quality speech.
- If higher number of parameters is used for quality purposes, the training becomes difficult due to the large number of parameters to be estimated.

Transmutative Voice Conversion

Transmutative Approach

- Use high-quality vocoders (like harmonic vocoder).
- They usually have high dimensions.
- Training a statistical transformation function is not robust.
- The goal is to modify high-dimensional features based on low-dimensional clues.

Transmutative Approach

- In this slide play the proof of concept of modifying harmonics based on formants
- Also play LSF VC files

Transmutative Approach

- Training
- we calculate the optimal parameter set

$$\lambda_{\mathcal{G}}^{*} = \arg\min_{\lambda_{\mathcal{G}}} \mathrm{E}_{\mathcal{G}} \left(\mathrm{Y}_{\uparrow}^{\mathrm{train}}, \mathcal{G}(\mathrm{X}_{\uparrow}^{\mathrm{train}}, \lambda_{\mathcal{G}})\right) \tag{3}$$

- where $\mathcal{G}$ is an operation that transmutes its input $\mathrm{X}_{\uparrow}^{\mathrm{train}}$ according to parameters$\lambda_{\mathcal{G}}$

Transmutative Approach

- Can also be calculated from a low dimensional feature set $X_{\downarrow}^{\text{train}}$

$$\lambda_{\mathcal{H}}^{*} = \arg\min_{\lambda_{\mathcal{H}}} \mathrm{E}_{\mathcal{H}} \left(\lambda_{\mathcal{G}}^{*}, \mathcal{H}(X_{\downarrow}^{\text{train}}, \lambda_{\mathcal{H}}) \right). \tag{4}$$

- This would be useful if we want the estimation be more robust, because the low-dimensional features have a smoother Error function.

Transmutative Approach

- Conversion:
  - find the best conversion function
  - manipulate the high-dimensional features based on the conversion function

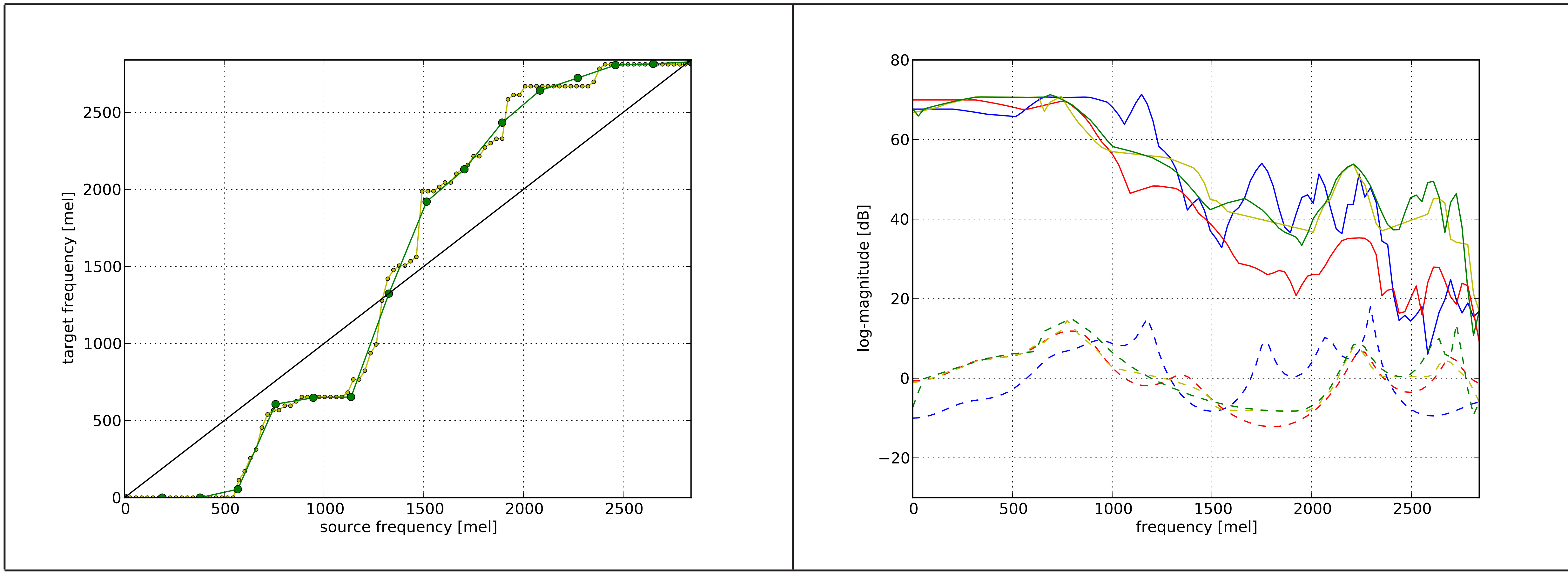
$$\mathcal{G}\left(\mathrm{X}_{\uparrow}^{\mathrm{test}}, \mathcal{H}(\mathrm{X}_{\downarrow}^{\mathrm{test}}, \lambda_{\mathcal{H}}^{*})\right) = \hat{\mathrm{Y}}_{\uparrow}^{\mathrm{test}}. \tag{5}$$

Transmutative Voice Conversion

- Proposed transmutation algorithm:
  - $\mathcal{G}$: spectral warping and amplification
  - $X_{\uparrow}$: source sinusoidal parameters
  - $Y_{\uparrow}$: target sinusoidal parameters
  - $\mathcal{H}$: A probabilistic, piece-wise linear mapping function

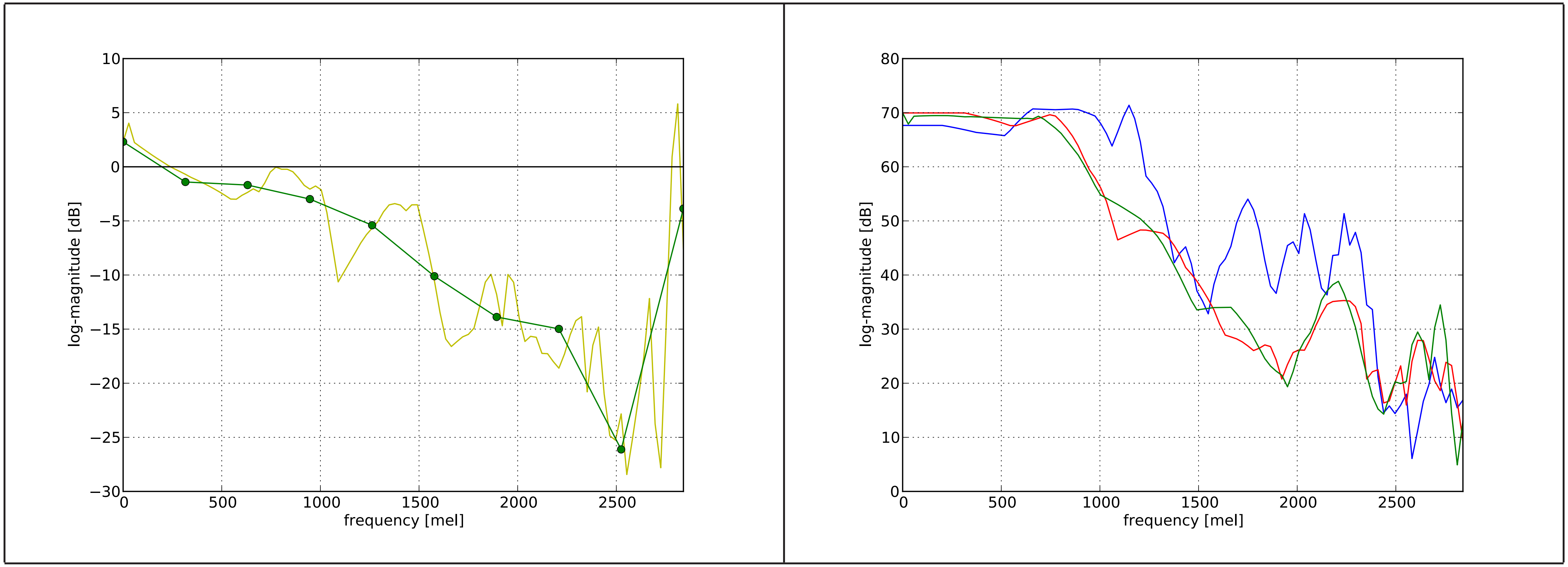
$$\lambda_{\mathcal{H}}^{*} = \arg\min_{\lambda_{\mathcal{H}}} \mathrm{E}_{\mathcal{H}}\left(\lambda_{\mathcal{G}}^{*}, \mathcal{H}(\mathrm{X}_{\downarrow}^{\mathrm{train}}, \lambda_{\mathcal{H}})\right). \tag{6}$$

Transmutative Voice Conversion



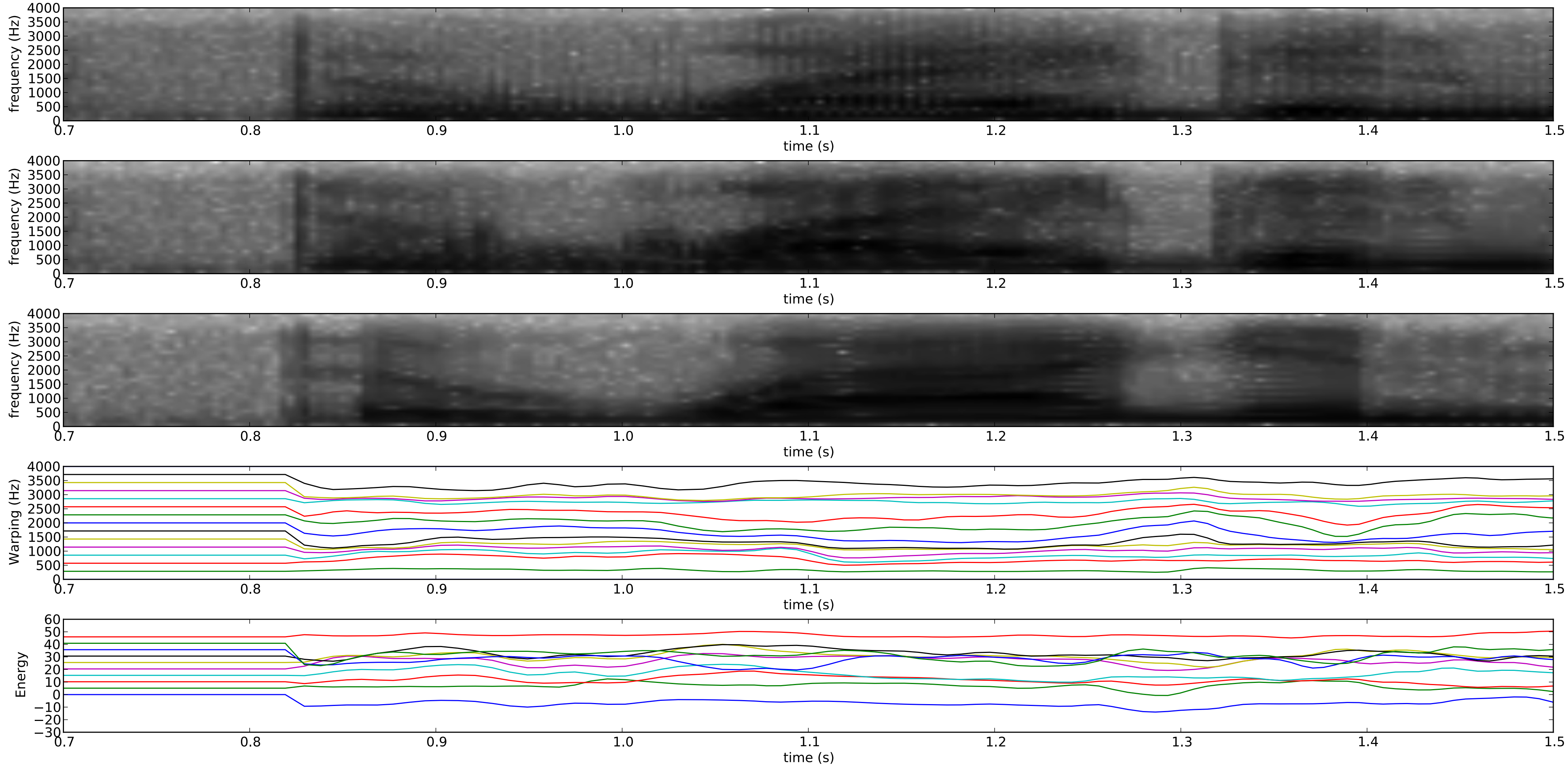
Source (blue) and target (red) magnitude spectra (solid lines), and their corresponding LPC spectra (blue and red dashed lines). Yellow lines are the result of applying the full (yellow) or parameterized warping function (green) to the source LPC (dashed) and original (solid) spectra.

Transmutative Voice Conversion



(left) Gain function (yellow) and its piece-wise linear parametrization (green) using 10 “knots” (green circles). The zero-gain line (black) is added for reference. (right) Source (blue), target (red), and warped and amplified source (green) magnitude spectra.

Transmutative Voice Conversion



Source (first panel), conversion (second panel), and target (third panel, time-aligned to the source for comparison purposes only) spectrograms, as well as corresponding warping (fourth panel) and gain (added by an arbitrary value for visualization) (fifth panel) parameter trajectories for the LPC-based conversion, for the utterance “mesh wire”.

Transmutative Voice Conversion

- These methods usually do not have a high quality because they are transformed to a lower dimension domain.
- Re-synthesis from them results in low quality speech.
- If higher number of parameters is used for quality purposes, the training becomes difficult due to the large number of parameters to be estimated.

Transmutative Voice Conversion

- play generative VC
- play transmutative VC

Evaluation

- We used 70 Harvard sentences
- Spoken by two male (M1, M2) and two female speakers (F1, F2)
- Sampling rate of 16 kHz
- Conversions:
  - Two cross-gender (M1→F1, F2→M2) and
  - Two intra-gender (M2→M1, F1→F2)
- # of training sentences: 46
- # of development sentences: 4
- # of test sentences: 20

Evaluation

- Generative VC (GEN) method:
  - Vocoder: LPC vocoder
  - Mapping: JDGMM with 16 mixture components
  - Features: 18 LSF coefficients
- Transmutative VC (FOR) method:
  - Vocoder: Harmonic vocoder
  - Modifications: Frequency warping using Formants + Gain manipulation
  - Estimating Modification Parameters: JDGMM with 3 mixture components from 18 LSFs to find estimate the warping (target formant locations) and gain parameters
- Transmutative VC (DFW) method:
  - Vocoder: Harmonic vocoder
  - Modifications: Frequency warping using Arbitrary Warping Function + Gain manipulation
  - Estimating Modification Parameters: JDGMM with 3 mixture components from 18 LSFs to find estimate the warping and gain parameters

- In this test, listeners hear two utterances A and B with different content
- They are asked to indicate wether they thought that A and B were spoken by the same or by two different speakers,
- They use five-point scale: +2 (definitely same), +1 (probably same), 0 (unsure), -1 (probably different), and -2 (definitely different)
- We considered the following five stimulus pairs: NAT-NAT, NAT-DFW, NAT-FOR, NAT-GEN, and NAT-LSF.

Evaluation

Speaker Recognizability

- Experiment is done is Amazon Mechanical Turk.
- 44 Listeners, approval rating > 90%,
- Each listener judged 40 sentence pairs, 10 trials for each of the four conversions. During these 10 trials, 2 trials were used for each of the 5 conditions.
- Sentence pairs were either:
  - "same": The "same" speaker (the conversion and the target),
  - "diff": The source speaker for intra-gender conversions, and the alternate speaker of the same gender for cross-gender conversion.

Evaluation

Speaker Recognizability

NAT-	NAT	DFW	FOR	GEN	LSF
same	1.39	-0.37	-0.38	0.12	1.04
diff	-1.32	-0.29	-0.68	-0.22	-1.08
all	1.36	-0.039	0.14	0.17	1.06

Table: Average speaker recognition test results (standard deviation is between 1.0 to 1.2) for diff, same and all conditions.

Evaluation

Speech Quality

- To evaluate conversion speech quality, we conducted a comparative mean opinion score (CMOS) test.
- Listeners hear two utterances A and B two different conditions,
- Indicate if B was better or worse than A, using a five-point scale consisting of +2 (much better), +1 (somewhat better), 0 (same), -1 (somewhat worse), -2 (much worse).
- We considered the following four stimulus pairs: FOR-GEN, DFW-GEN, GEN-NAT, FOR-NAT.

Speech Quality

	FOR-GEN	DFW-GEN	GEN-NAT	FOR-NAT
all	-0.43(1.4)	0.88(0.9)	1.83(0.4)	1.57(1.1)

Table: Average preference test results (standard deviation in parentheses).

Discussion

- We can achieve higher quality speech using Transmutative VC.
- When the modification function is not smooth through time, the quality is very poor specially because of sudden changes to the spectrum.
- The degree of change does not seem to be adequate, because the similarity scores are lower than generative method.

Conclusion

- conclusion