Transmutative Voice Conversion

ICASSP 2013

S. H. Mohammadi and A. Kain

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

May 31, 2013

Voice Conversion (VC)

process the speech of a *source* speaker to sound like a *target* speaker

- Applications
  - personalized TTS
    - for individuals with disabilities
    - message readers with custom/sender identities
  - movie dubbing
  - interpretive services by human or machine
- Important criteria
  - speaker recognizability
  - speech quality

Categorize approaches as:

1. Generative
   - compact parametrization of speech
   - direct mapping from input to output parameters
   - quality is limited by parametric vocoder
2. Transmutative
   - high-fidelity speech model
   - difficult to train satisfactory direct mapping
     - high-dimensional feature space
     - especially for very small training sets
   - instead, use a constrained mapping
     - change prominent characteristics while leaving less-prominent characteristics unmodified

Generative approach

1. Obtain aligned source and target feature sequences $X_{N\times d}$ and $Y_{N\times d}$
   - align parallel data using DTW and/or phonetic boundaries + frame interpolation
   - non-parallel data approach [e. g. Mouchtaris 2004, Erro 2010, Godoy 2012, Song 2013]
2. Training:
$$\lambda^* = \arg\min_{\lambda} E\left(Y^{train}, \mathcal{F}(X^{train}; \lambda)\right)$$
3. Conversion:
$$\mathcal{F}(X^{test}; \lambda^*) = \hat{Y}^{test}$$

Generative approach — examples

- Implementations of $\mathcal{F}$:
  - VQ [Abe 1988]
  - ANN [Narendranath 1995]
  - GMM [Stylianou 1998, Toda 2007, Zen 2011]
  - SVM [Laskar 2009]
- Speech model / feature types:
  - formants [Narendranath 1995]
  - line spectral frequencies (LSF) [Kain 1998, Arslan 1999]
  - cepstrum [Toda 2001]

Problem

Speech quality is limited by the parametric vocoder

Transmutative approach

- Consider high-dimensional features $X_{\uparrow}^{train}$ and $Y_{\uparrow}^{train}$ $(d > 100)$
- Training a direct mapping is difficult, instead:

$$\lambda_{\mathcal{G}}^{*} = \arg\min_{\lambda_{\mathcal{G}}} E_{\mathcal{G}}\left(Y_{\uparrow}^{train}, \mathcal{G}(X_{\uparrow}^{train}; \lambda_{\mathcal{G}})\right)$$

  - $\mathcal{G}$ "transmutes" its input $X_{\uparrow}^{train}$ according to parameters $\lambda_{\mathcal{G}}$
  - $\lambda_{\mathcal{G}}^{*}$ is relatively easy to compute

Key concept

$\mathcal{G}$ is constrained in such a way that its possible outputs are congruent with the types of changes one expects *a priori* when converting one voice into another

Transmutative approach

- During conversion, need to predict transmutation parameters $\lambda^*_{\mathcal{G}}$
- Therefore, we train a mapping function such that

$$\mathcal{H}(X_{\downarrow}^{train}; \lambda^*_{\mathcal{H}}) = \hat{\lambda}^*_{\mathcal{G}}$$

  - note that this uses low-dimensional features $X_{\downarrow}^{train}$

- Conversion:

$$\mathcal{G}\left(X_{\uparrow}^{test}, \mathcal{H}(X_{\downarrow}^{test}, \lambda^*_{\mathcal{H}})\right) = \hat{Y}_{\uparrow}^{test}$$

Transmutative approach — examples

- Implementations of $\mathcal{G}$ (frame-based):
  - frequency warping
    - unconstrained DFW [Valbret 1992]
    - VTLN-based DFW [Sündermann 2003, Erro 2012]
    - use formant frequencies to guide frequency warping [Erro 2007]
  - additional low-order gain adjustment [Erro 2007, Godoy 2012]
- Speech model / feature types:
  - Fourier transform [Sündermann 2004]
  - high-order (discrete, warped) cepstrum [Erro 2007, Godoy 2012]

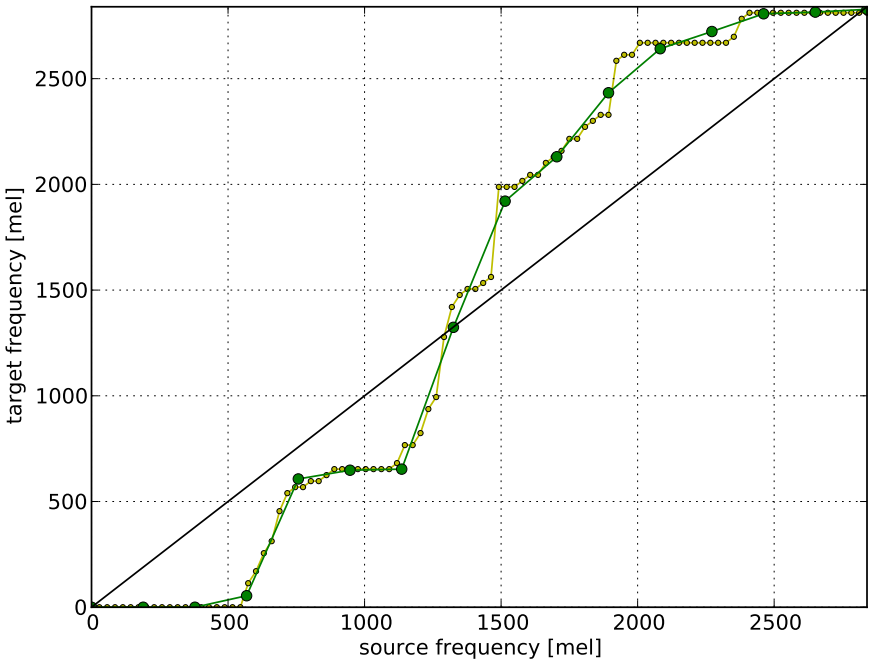
Proposed method

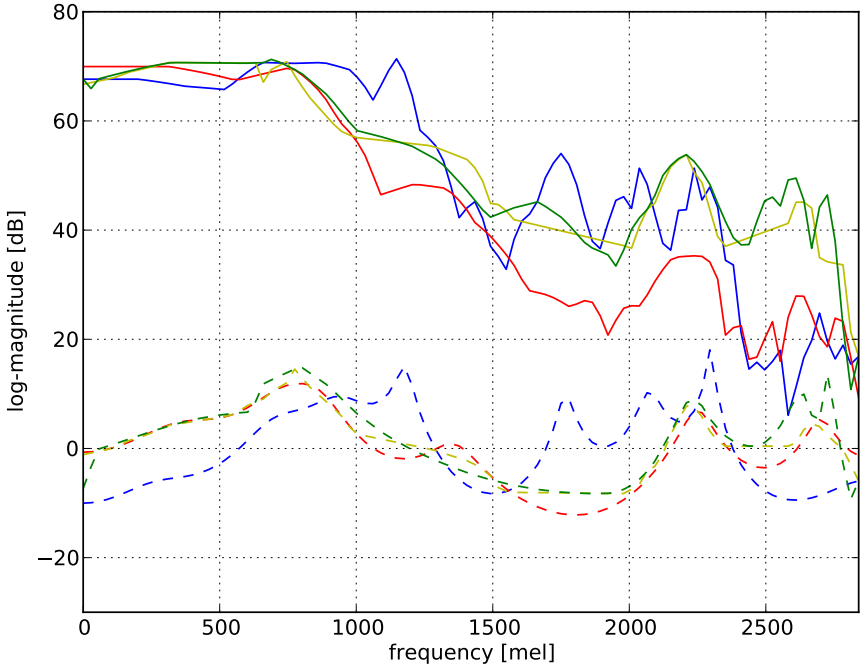
- Transmutative conversion:

$$\mathcal{G}\left(X_{\uparrow};\lambda_{\mathcal{G}}\right)=\hat{Y}_{\uparrow},\quad \lambda_{\mathcal{G}}=\mathcal{H}(X_{\downarrow};\lambda^{*}_{\mathcal{H}})$$

- $\mathcal{G}$: frame-based spectral warping and gain function *using piece-wise linear parametrization*
  1. based on DFW of LPC spectrum (DFW), $X_{\downarrow}$: low-order cepstral features
  2. based on formant frequencies + *bandwidth preservation* (FOR), $X_{\downarrow}$: formant frequencies
- $X_{\uparrow}$ and $Y_{\uparrow}$: source/target magnitude spectra of a harmonic model
  - pitch-synchronous over 2 frames, warped via mel-scale, resampled to 100 points using spline interpolation
- $\mathcal{H}$: conventional joint-density GMM

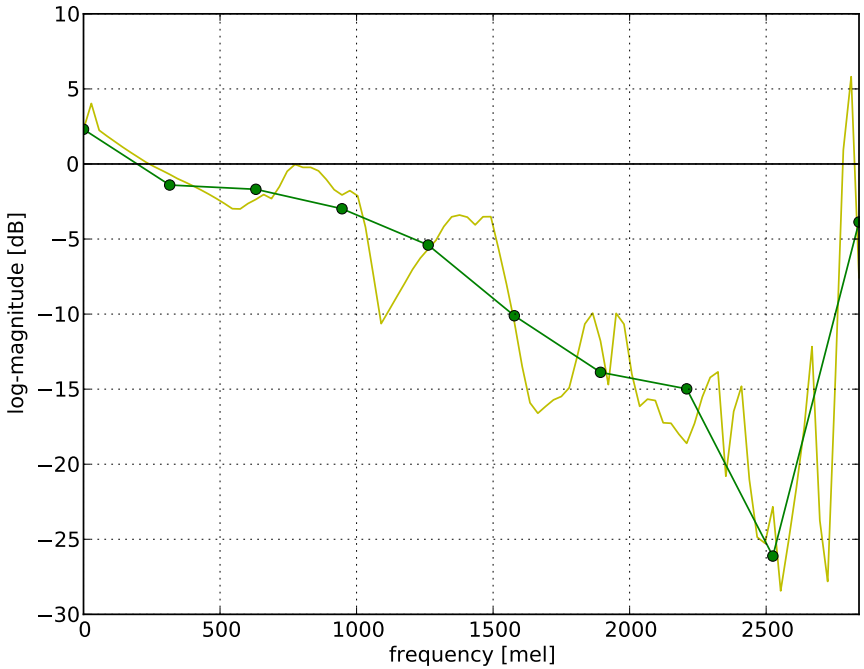
Piece-wise linear DFW based on LPC (DFW)

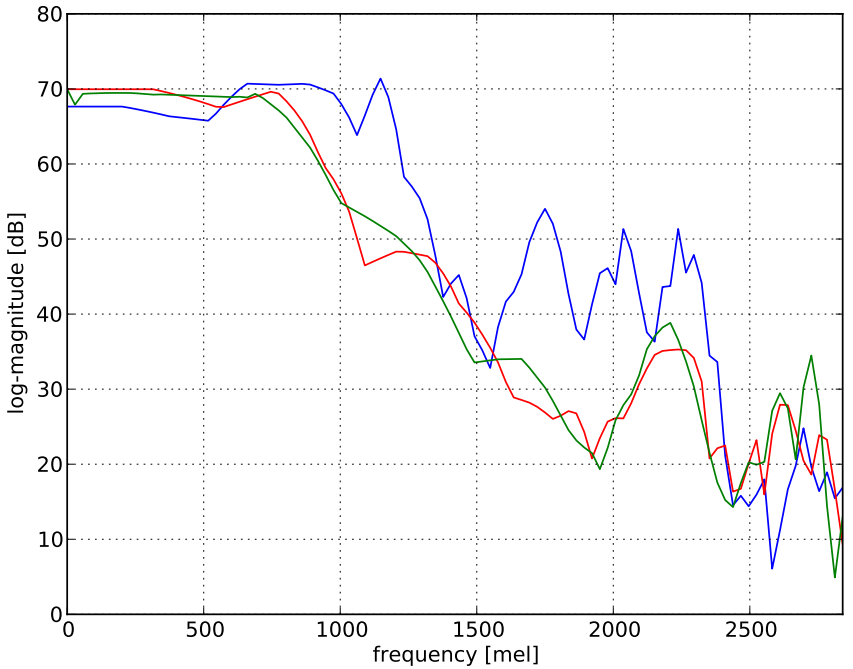




Source (blue) and target (red) magnitude spectra (solid lines), and their corresponding flattened LPC spectra (blue and red dashed lines). Yellow lines are the result of applying the full (yellow) or parameterized warping function (green) to the source LPC (dashed) and original (solid) spectra.

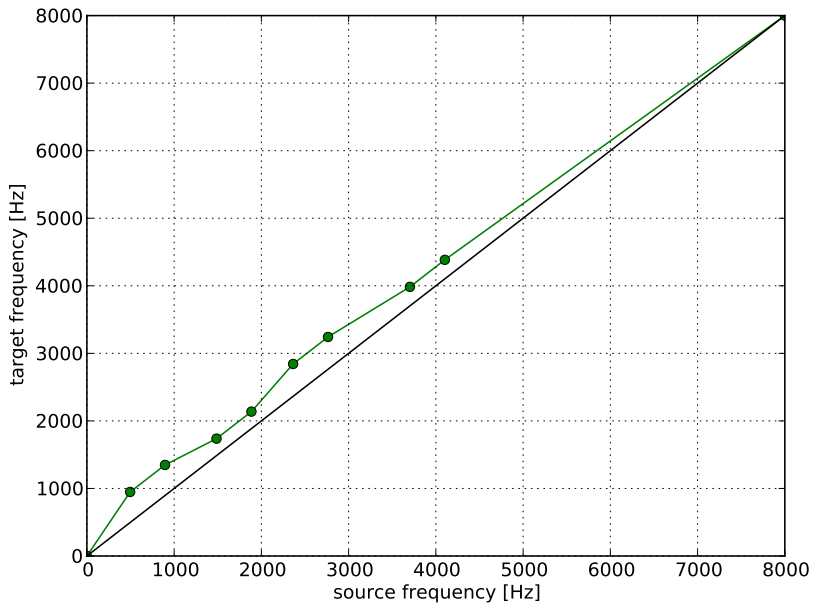
Piece-wise linear gain function





Gain function (yellow) and its parametrization (green). Source (blue), target (red), and warped and amplified source (green) magnitude spectra.

Piece-wise warping based on formants (FOR)



- x and y coordinate of parameterized warping function defined by source and target formant frequencies
- In addition, we subtract and add a constant 150 Hz to both the x- and the y-coordinate of each of the four knots
  - special handling where these frequencies overlap
  - effect is bandwidth preservation

Setup

- Used 70 Harvard sentences:
  - 46 training sentences
  - 4 development sentences
  - 20 of test sentences
- 4 speakers: two male (M1, M2) and two female (F1, F2)
- 4 conversion pairs
  - 2 cross-gender (M1→F1, F2→M2)
  - 2 intra-gender (M2→M1, F1→F2)
- AMT listeners
  - approval ratings of at least 90%
  - located in North America

Stimulus conditions

NAT	Natural waveform
LPC	LPC vocoder re-synthesis
GEN	Generative approach (JDGMM-16, LSF-18)
DFW	Transmutative approach with DFW based on LPC
FOR	Transmutative approach with FW based on formants

- Conversion stimuli also feature basic prosodic conversion
  - matched target F0 mean and variance
  - matched target speaking rate
- Loudness-normalized using an A-weighted RMS energy measure

Speaker recognizability

- Listeners hear two utterances with *different* content, where one is a conversion speaker and the other speaker is:

  same the target of the conversion, e. g. (M1→F2, F2)
  diff a *same-gender* alternate speaker, e. g. (M1→F2, F1)

- Five condition pairs:
  1. NAT-NAT (human performance as baseline)
  2. NAT-LPC (effect of parametric vocoder, max performance of GEN)
  3. NAT-DFW
  4. NAT-FOR
  5. NAT-GEN

- 44 listeners rated 40 sentence pairs' speaker similarity on scale

Speaker recognizability — results

NAT-	NAT	LPC	DFW	FOR	GEN
same	1.39	1.04	-0.37	-0.38	**0.12**
diff	-1.32	-1.08	-0.29	**-0.68**	-0.22

- -2 definitely different, -1 probably different, 0 unsure, +1 probably same, +2 definitely same
- No statistically significant differences between (NAT-FOR, NAT-GEN) or (NAT-FOR, NAT-DFW)

Speech quality

- Standard comparative mean opinion score (CMOS) test using two utterances with
  - *same* content and *same* speakers
  - two *different* conversion approaches
- 4 condition pairs:
  1. FOR-GEN
  2. DFW-GEN
  3. GEN-NAT
  4. FOR-NAT
- 35 listeners rated 40 sentence pairs' relative speech quality on scale

Speech quality — results

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

- +2 (much better), +1 (somewhat better), 0 (same), -1 (somewhat worse), -2 (much worse)
- A two-tailed *t*-test shows statistical significance between FOR-NAT and GEN-NAT ($t(188) = 2.24$, $p = 0.026$)

Conclusion

- Speaker recognizability:
  - FOR obtained good scores for making a speaker dissimilar
  - GEN was still better in creating a specific speaker
    - $\mathcal{G}$ may still be too simple
- Speech quality:
  - results suggest FOR $>$ GEN $>$ DFW
- Piece-wise linear parametrization of DFW was not successful
  - likely because sometimes the raw DFW resulted in bad alignments
    - may explain success of simpler models such as the bilinear transform

Open question

How do we reliably align two spectra with relatively flexible models?

Demo

1. Vocoder quality differences:
    1. original
    2. harmonic
    3. LSF
2. Voice conversion:
    1. target original
    2. source original
    3. source with prosody=target
    4. source with prosody=target, transmutative
    5. source with prosody=target, generative
    6. target original

Thank you!

Questions?