High Quality Voice Morphing

Hui YE
Cambridge University Engineering Department

March 16th 2004

Content

- Motivation
- Key issues of Voice Morphing
- Pitch Synchronous Harmonic Model
- Speaker Identity Feature
- Conversion Function
- System Enhancement
- Evaluation
- Unknown Speaker Transformation
- Summary

What is Voice Morphing?

Voice morphing is a technique for modifying a source speaker's speech to sound as if it was spoken by some designated target speaker.

- Research Goals: To develop algorithms which can morph speech from one speaker to another with the following properties.

  1. High quality (natural and intelligible)
  2. Morphing function can be trained automatically from speech data which may or may not require the same utterances to be spoken by the source and target speaker.
  3. the ability to operate with target voice training data ranging from a few seconds to tens of minutes.

Key Technical Issues

1. Mathematical Speech Model
   - For speech signal representation and modification
2. Accoustic Feature
   - For speaker identification
3. Conversion Function
   - Involves methods for training and application

Pitch Synchronous Harmonic Model

Sinusoidal model has been widely used for speech representation and modification in recent years.

1. PSHM is a simplification of the standard ABS/OLA sinusoidal model

$$\tilde{s}_k(n) = \sum_{l=0}^{L_k} A_l^k cos(l\omega_0^k n + \phi_l^k), [n = 0, \cdots, N_k] \tag{1}$$

2. The parameters were estimated by minimising the modeling error

$$\mathcal{E} = \sum_n [s_k(n) - \tilde{s}_k(n)]^2 \tag{2}$$

Time and Pitch Modification using PSHM

1. Pitch Modification

   - It is essential to maintain the spectral structure while altering the fundamental frequency.
   - Achieved by modifying the excitation components whilst keeping the original spectral envelope unaltered.

2. Time Modification

   - PSHM model allows the analysis frames be regarded as phase-independent units which can be arbitrarily discarded, copied and modified.

3. Demo original speech $\Longrightarrow$ pitch scale 1.3 $\Longrightarrow$ time scale 1.3

Speaker Identity Features

1. Suprasegmental Cues
    - Speaking rate, pitch contour, stress, accent, etc.
    - Very hard to model
2. Segmental Cues
    - Formant locations and bandwidths, spectral tilt, etc.
    - Can be modelled by spectral envelope.
    - In our research, Line Spectral Frequencies (LSF) are used to represent the spectral envelope.
    - LSF requires less coefficients to efficiently capture the formant structure and it has better interpolation properties.

Conversion Function

- Interpolated Linear Transform
  - Assume the training set contains two sets of time-aligned parallel spectral vectors (LSF) $\mathbf{X}$ and $\mathbf{Y}$ respectively from the source and target speaker.

$$\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_T}] \ \mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_T}] \tag{3}$$

  - Use a GMM to cluster the source speech data into M classes, and each class is associated with a linear transform.
  - The conversion function is defined as

$$\mathcal{F}(\mathbf{x}) = (\sum_{m=1}^{M} \lambda_m(\mathbf{x}) W_m)\bar{\mathbf{x}} \quad where \quad \lambda_m(\mathbf{x}) = P(C_m|\mathbf{x}) \tag{4}$$

Conversion Function (Cont)

- For easier manipulation, equation (4) can be rewritten compactly as,

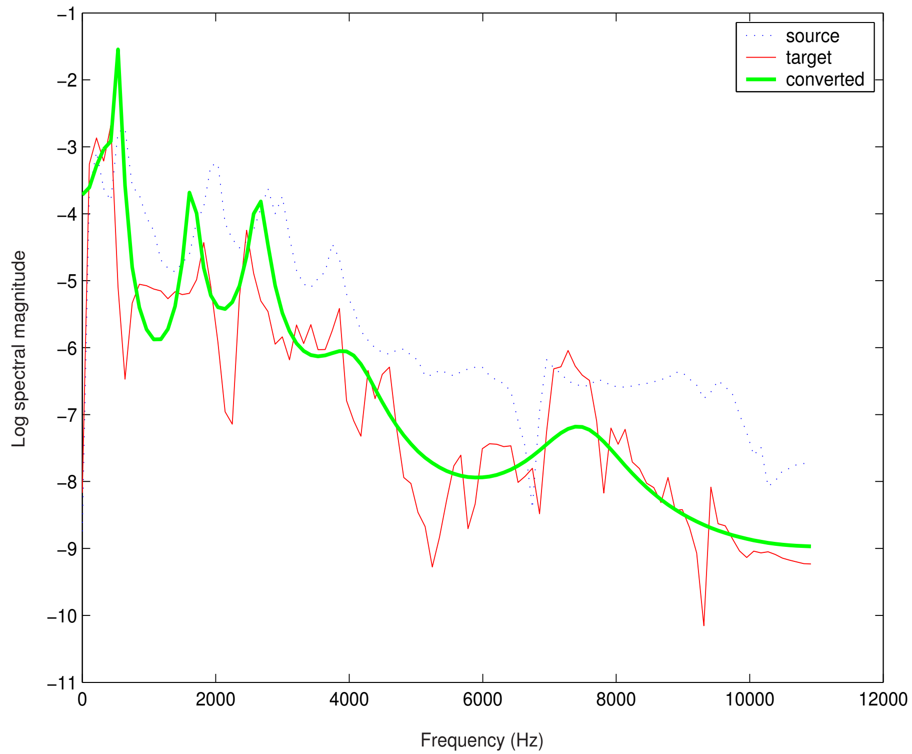
$$
\begin{aligned}
\mathcal{F}(\mathbf{x}) &= \left[W_1 \vdots W_2 \vdots \cdots \vdots W_M\right] \begin{pmatrix} \lambda_1(\bar{\mathbf{x}})\bar{\mathbf{x}} \\ \cdots \\ \lambda_2(\bar{\mathbf{x}})\bar{\mathbf{x}} \\ \cdots \\ \vdots \\ \cdots \\ \lambda_M(\bar{\mathbf{x}})\bar{\mathbf{x}} \end{pmatrix} \\
&= \bar{\mathbf{W}}\Lambda(\mathbf{x}) \qquad (5)
\end{aligned}
$$

- Use standard least-squares estimation to derive the transformation matrices $\bar{\mathbf{W}}$.

$$\bar{\mathbf{W}} = \mathbf{Y}\Lambda(\mathbf{X})'\Big(\Lambda(\mathbf{X})\Lambda(\mathbf{X})'\Big)^{-1} \qquad (6)$$

System Enhancements

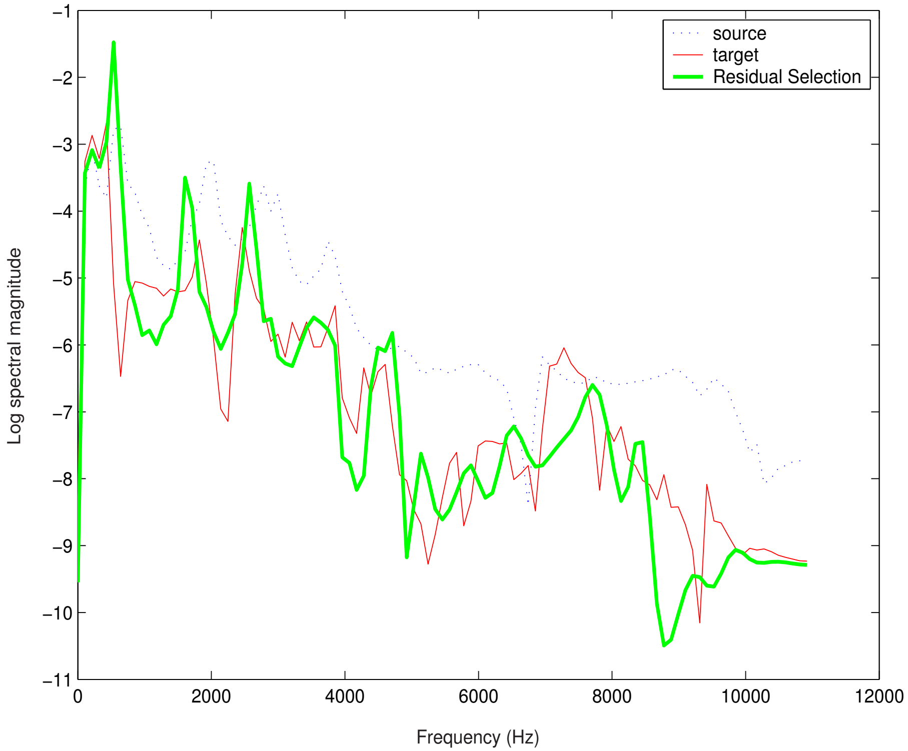
- Spectral Distortion
  - Formant structure has been transformed
  - Spectral details lost due to reduced LSF dimensionality
  - Spectral peaks broadened by the averaging effect of least square error estimation.



System Enhancements (Cont)

- Spectral Residual Selection
  - Idea: reintroduce the lost spectral details to the converted envelopes.
  - Use a codebook selection method to construct a residual.
  - Each target spectral residual r_t is associated with a LSF vector v_k.
  - The residual whose associated v_k minimizes the following square error is then selected.

$$\mathcal{E} = (v_k - \tilde{v})'(v_k - \tilde{v}) \qquad (7)$$



Unnatural Phase Dispersion

- In the baseline system, the converted spectral envelope was combined with the original phases. This results in converted speech with a "harsh" quality.

- Spectral magnitudes and phases of human speech are highly correlated.

- To simultaneously model the magnitudes and phases and then convert them both via a single unified transform is extremely difficult.

Phase Prediction

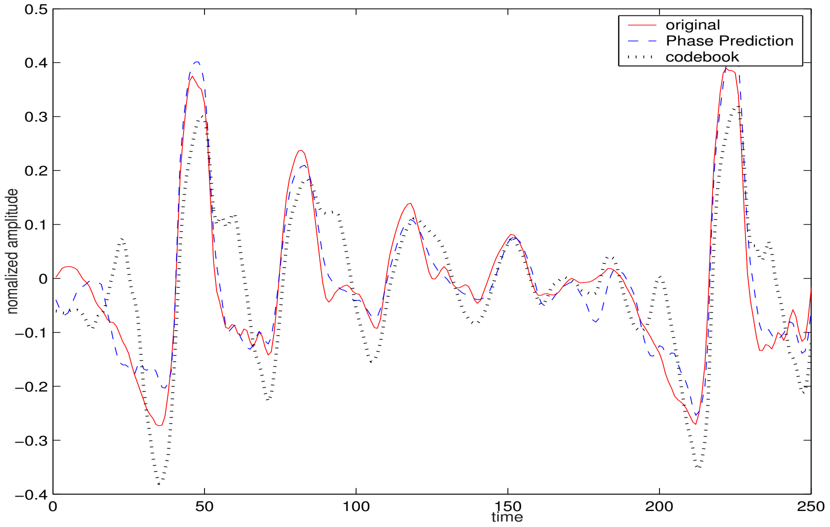
- Idea: Phase dispersion can be obtained from the waveform shape. If we can predict the waveform shape, then we can predict the phases.
- Implementation
  - A GMM model is first trained to cluster the target spectral envelopes coded via LSF coefficients into M classes $(C_1, \cdots, C_M)$.
  - For each target envelope v we have a set of posterior probabilities $\mathcal{P}(v) = [P(C_1|v), \cdots, P(C_M|v)]'$, this can be regarded as another form of representation of the spectral shape.
  - A set of template signal (codebook entries) $\mathcal{T} = [T_1, \cdots, T_M]$ can be estimated by minimising the waveform shape prediction error

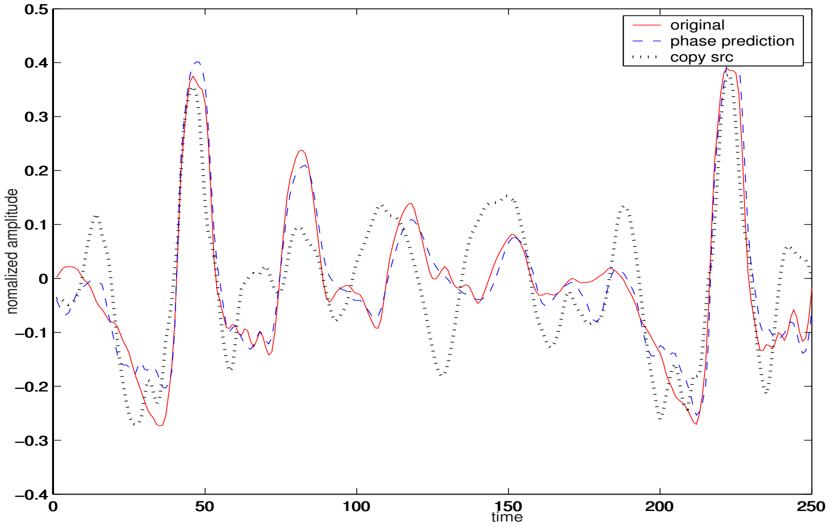
$$E = \sum_{t=1}^{N} (s(t) - \mathcal{T}\mathcal{P}(v_t))'(s(t) - \mathcal{T}\mathcal{P}(v_t)) \qquad (8)$$

Phase Prediction Result

The SNR ratio in dB of different phase coding methods.

src phases	codebook phases	phase prediction
3.2171	6.1544	7.2079





Transforming Unvoiced Sounds

- Theoretically the unvoiced sounds contain very little vocal tract information, so in our baseline system, the unvoiced sounds are not transformed.

- In reality, many unvoiced sounds have some vocal tract colouring which affects the speech characteristics.

- Since the spectral envelopes of the unvoiced sounds have large variations, it is not effective to convert them using the linear transformation scheme.

- An simple approach based on unit selection and concatenation was therefore developed to transform the unvoiced sounds.

Post-filtering

- As noted earlier, transform-based voice conversion has a tendency to broaden the formants.

- To mitigate this effect and suppress noise in the spectral valleys, a final post-processing stage applies a perceptual filter to the converted spectral envelope.

$$H(\omega) = \frac{A(z/\beta)}{A(z/\gamma)}, 0 < \gamma < \beta \leq 1 \tag{9}$$

where $A(z)$ is the LPC filter and the choice of parameters in our system is $\beta = 1.0$ and $\gamma = 0.94$.

Objective Evaluation

- Objective Measure: log spectral distortion defined as

$$d(S_1, S_2) = \sum_{k=1}^{L} (log a_k^1 - log a_k^2)^2 \tag{10}$$

where $\{a_k\}$ are the amplitudes resampled from the spectral envelope S at L uniformly spreaded frequencies.

- The overall transformation performance can be evaluated by comparing the converted-to-target distortion with the source-to-target distortion, which was defined as,

$$D = 10 log_{10} \frac{\sum_{t=1}^{N} d(S_{tgt}(t), S_{cov}(t))}{\sum_{t=1}^{N} d(S_{tgt}(t), S_{src}(t))} \tag{11}$$

where $S_{tgt}(t)$,$S_{src}(t)$ and $S_{cov}(t)$ are the target spectral envelope,

Experiments

- Training data: parallel speech data about 5 minutes from both the source and the target speaker.
- Tasks: 4 tasks; male to male, male to female, female to male and female to female.

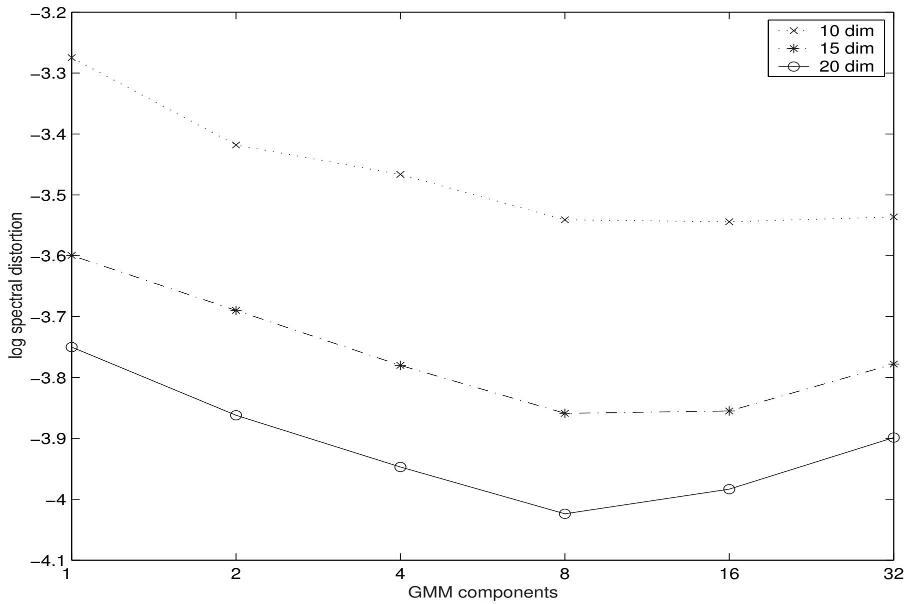


Fig: Log spectral distortion over different GMM components and dimensions.

Subjective Evaluation

- ABX test

	baseline system	enhanced system
ABX	86.4%	91.8%

- Preference test

	baseline system	enhanced system
preference	38.9%	61.1%

Examples

Voice transformatin with parallel training data.

	Source	Target	Converted 1	Converted 2
F to M	src01	tgt01	vc01	vm01
M to F	src02	tgt02	vc02	vm02
F to F	src03	tgt03	vc03	vm03
M to M	src04	tgt04	vc04	vm04

Unknown Speaker Voice Transformation

- Situation: No pre-existed training data is available from the source speaker, although there is still a reasonable amount of speech data from the designated target speaker.
- Idea: Use speech recognition to create a mapping between the unknown input source speech and the target vectors.
- Implementation:
  1. Use speaker independent HMM to force align the target data, then label each frame with a state id.
  2. Either force align or recognize the input source utterance, and label each frame with a state id.
  3. According to the force aligned or recognized state sequence, select the best matched target frames, and then train the transform.
  4. Apply the transform to the source utterance.

Examples

	Source	Converted	Target
Female	src05	vc05	tgt05
Male	src06	vc06	tgt06

Summary

- A complete solution to the voice morphing problem has been developed which can deliver reasonable quality.
- A system for unknown speaker transformation has been implemented which only requires pre-existed training data for the target speaker.
- However, there still some way to go before these techniques can support high fidelity studio applications.
- Future work would be
  - Improve the quality of the converted speech.
  - Cross language voice conversion will be another challenge.