

L20: voice conversion

Introduction

Codebook-based voice conversion

GMM-based voice conversion

Frequency-warping-based voiced conversion

Voice conversion w/o parallel corpus

Foreign accent conversion

Introduction

Voice transformation

- Aimed at controlling non-linguistic information in the speech signal, such as voice quality and voice individuality
- Voice transformations can be applied to the source and to the filter
 - Source transformations are generally restricted to prosodic modification; these techniques were reviewed in the previous lecture
 - Various methods have also been proposed to transform the glottal transfer function; see Childers (1995)
 - Transformations of the vocal tract filter, generally performed at the segmental level, are the subject of this lecture
 - Global transformations of the filter can also be used, e.g., through VTLN (as in speaker normalization for LVCSR), but the conversions are limited

Terminology

- Voice transformation
 - Seeks to transform filter characteristics in a general way, without reference to a target speaker
- Voice conversion
 - Seeks to transform the filter of a speaker (the source speaker) to have similar characteristics to that of another speaker (the target speaker)
- Voice morphing
 - Seeks to blend two source voices to form a third voice that contains characteristics of the two sources
- Accent conversion
 - Seeks to transform the regional/foreign accent of a source speaker to have similar characteristics to that of another accent (e.g., native)
- Cross-language conversion
 - Seeks to translate utterances spoken in one language into the equivalent ones of another language (e.g., through ASR→MT→TTS)

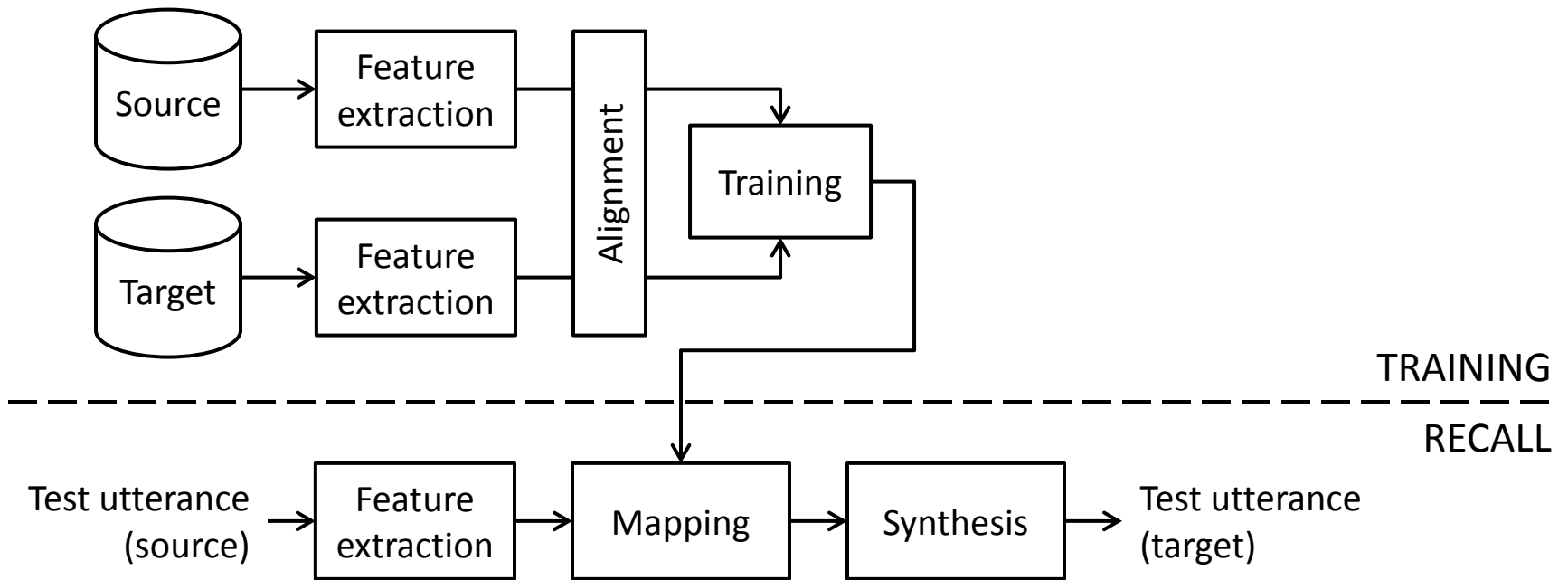
Applications of voice conversion

- Customization of text-to-speech systems
 - VC requires considerably less training data (10 min) than that required for building a TTS from scratch (10 hours)
- Voice editing and dubbing
 - e.g., regenerating voices of actors/actresses who are no longer alive or who've lost their voice to old age or illness
- Personalized human computer interfaces
 - Generating a virtual copy of voices that are meaningful to the user, e.g., a mother's voice in educational software
- Medical applications
 - Voice restoration systems
 - Training interfaces for speech pathologies
- Other applications
 - Entertainment, special effects, gaming
 - Testing speaker recognition systems
 - Improved ASR

Types of voice conversion techniques

- According to the mapping
 - Vector Quantization
 - Gaussian Mixture Models
 - Dynamic Frequency Warping
 - Combination of DFW with GMM or VQ
- According to the corpus
 - Based on parallel recordings
 - Based on non-parallel recordings

Building blocks of a voice conversion system



Codebook-based voice conversion

Overview

- One of the first methods for voice conversions, proposed by Abe et al. (1988) based on a speaker adaptation method by Shikano et al. (1986)
- Method consists of builds a discrete mapping between source and target spectral envelopes

Approach (Abe et al., 1988)

- Assume a corpus of parallel recordings from two speakers (A and B)
- Generate spectral vectors x_t and y_t (e.g., MFCC) for all utterances
- Learn a vector quantization model for each speaker
- For each utterance, align frames for the two speakers using DTW
- Store vector correspondences b/w two speakers as a 2D histogram
- Using each histogram as a weighting function, define the mapping codebook as a linear combination of speaker's B vectors
- Mappings for pitch and energy are built simultaneously using a similar method

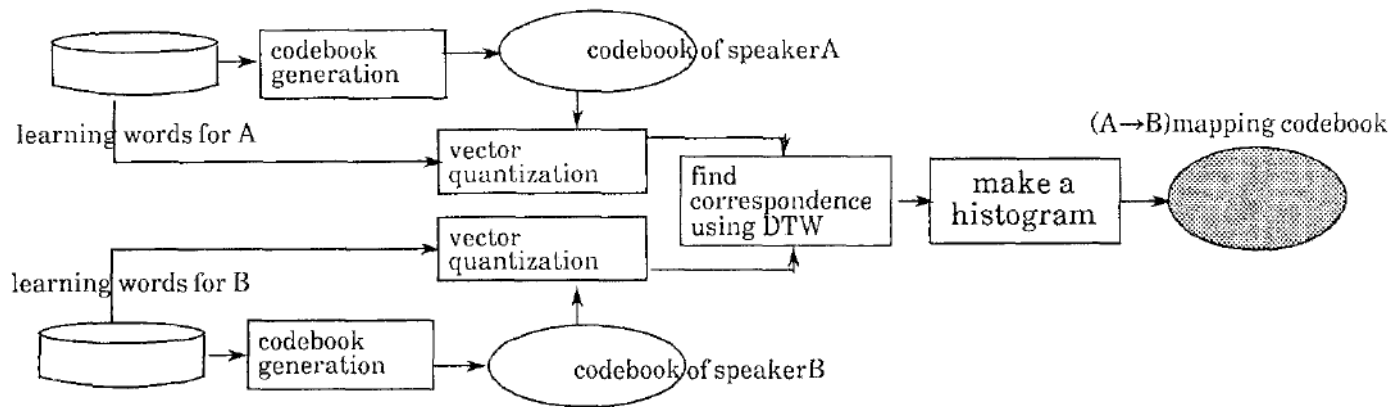


Fig.1. Method for Generating a Mapping Codebook

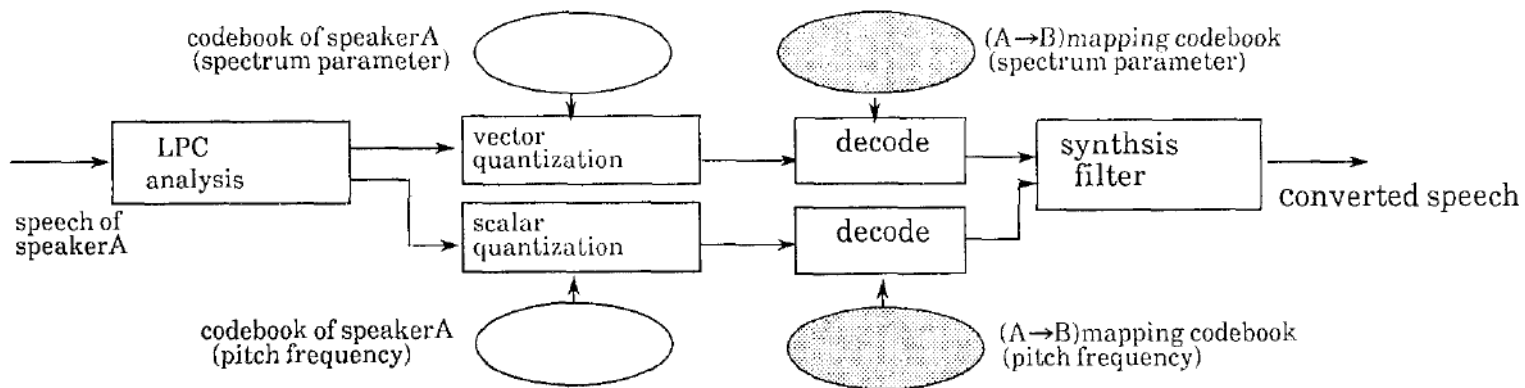


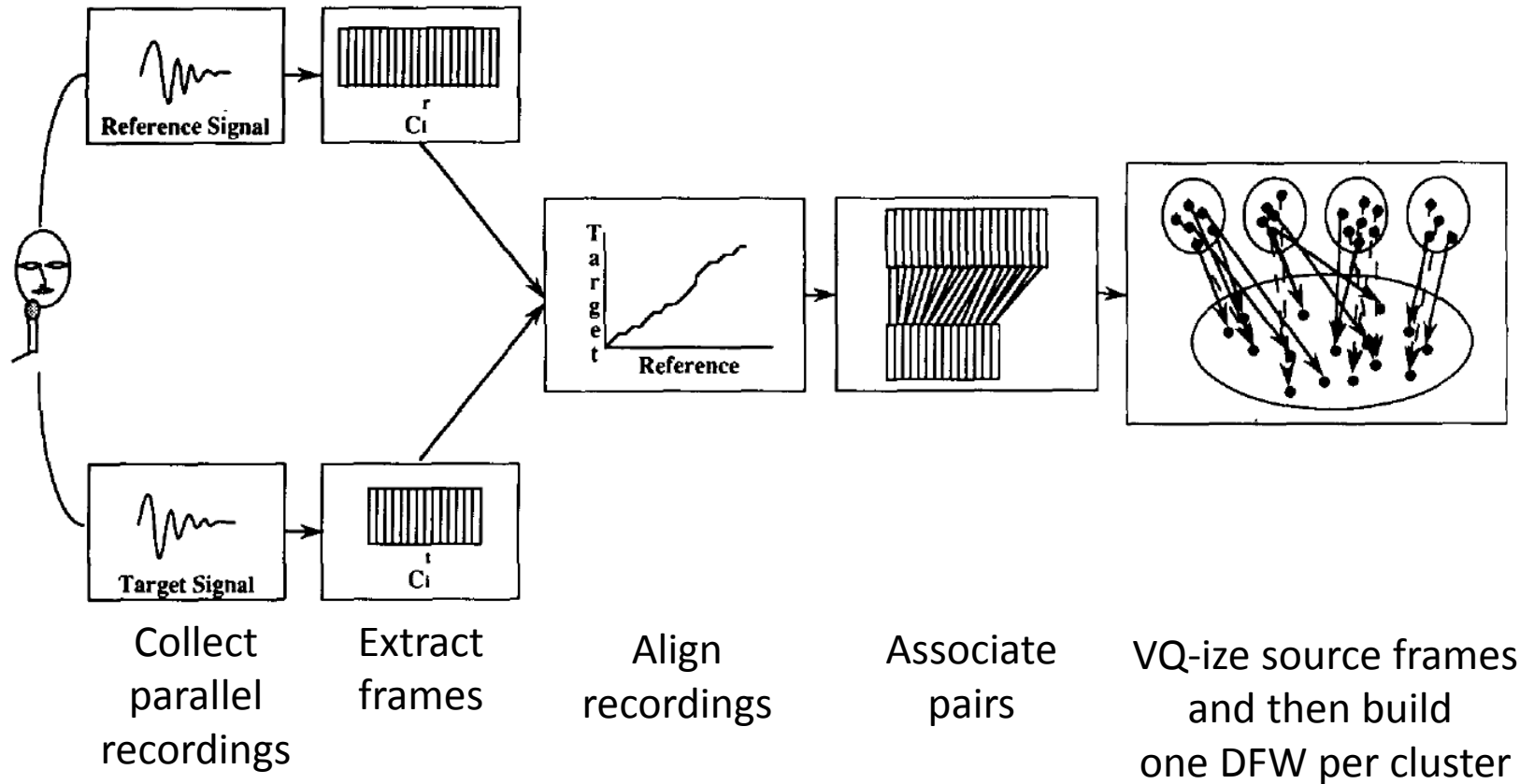
Fig.2. Block Diagram of a Conversion from Speaker A to Speaker B

[Abe et al., 1988]

Alternative approach (Valbret et al., 1992)

- Assume a corpus of parallel recordings from two speakers (A and B)
- Align recordings from the two speakers using DTW
- Perform source/filter decomposition through LP analysis in a pitch-synchronous fashion
- Vector quantize the distribution of spectral envelopes for the source speaker
- For each cluster, build a dynamic frequency warp (DFW) from source spectral envelope to target spectral envelope
 - The DFW of each cluster is defined as the median DFW of all pairs of frames in that cluster
- (Prosodic modification is performed through LP-PSOLA)

VC through frequency warping of codewords



[Valbret et al., 1992]

GMM-based voice conversion

Approach (Stylianou, 1998)

- Assume two sets of paired spectral vectors x_t and y_t corresponding to the spectral envelope (e.g., MFCC) of the source and target speakers
 - The two sets are of the same length and are assumed to be aligned
 - We seek to find a mapping function $y_t = \mathcal{F}(x_t)$
- First, build a GMM for the source data $\{x_t, t = 1 \dots n\}$

$$p(x) = \sum_{i=1}^m \alpha_i N(x; \mu_i, \Sigma_i)$$

- where

$$N(x; \mu, \sigma) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]; \quad s.t. \sum_{i=1}^m \alpha_i = 1$$

- In the case where $m = 1$ (single Gaussian) and both distributions are jointly Gaussian, it can be shown that the MMSE estimate is

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

- where $\nu = E[y]$ and $\Gamma = E[(y - \nu)(y - \nu)^T]$

- Thus, for the multimodal case ($m > 1$), we assume a similar form for the conversion function

$$\mathcal{F}(x) = \sum_{i=1}^m P(C_i|x_t) [\nu_i + \Gamma_i \Sigma_i^{-1}(x_t - \mu_i)]$$

- and seek parameter values that minimize the total squared error

$$\epsilon = \sum_{t=1}^n \|y_t - \mathcal{F}(x_t)\|^2$$

Special cases

– Full conversion

- The general case, when the GMM has full covariance matrices

– Diagonal conversion

- Covariance matrices $\{\Gamma_i, \Sigma_i\}$ are assumed to be diagonal, which is reasonable for cepstral vectors since they are approximately orthogonal
- This significantly reduces computational load because (i) it requires fewer parameters and (ii) the system of equations becomes decoupled

– VQ-type conversion

- If we omit the correction term $\Gamma_i \Sigma_i^{-1} (x_t - \mu_i)$, the conversion becomes

$$\mathcal{F}(x) = \sum_{i=1}^m P(C_i | x_t) v_i$$

- which is a form of weighted codebook mapping
- Note that this form of conversion is severely restricted as it only allows target spectral envelopes that are linear combinations of codewords v_i

Optimization (full conversion)

- The solution to the conversion function can be expressed as

$$\mathbf{y} = \mathbf{P} \cdot \mathbf{v} + \mathbf{\Delta} \cdot \mathbf{\Gamma} = [\mathbf{P} \ : \ \mathbf{\Delta}] \cdot \begin{bmatrix} \mathbf{v} \\ \mathbf{\Gamma} \end{bmatrix}$$

- where \mathbf{y} is a $n \times p$ matrix containing the target vectors
- \mathbf{P} is a $n \times m$ matrix containing the conditional probabilities $P(C_i|x_t)$, which can be computed from the GMM using Bayes rule
- $\mathbf{v} = [v_1 v_2 \dots v_m]^T_{(m \times p)}$ and $\mathbf{\Gamma} = [\Gamma_1 \Gamma_2 \dots \Gamma_m]^T_{((m \times p) \times p)}$ are the unknown parameters of the conversion function, and
- $\mathbf{\Delta}$ is a $n \times pm$ matrix defined by blocks as

$$\mathbf{\Delta} = \begin{bmatrix} p_1(1)(x_1 - \mu_1)^T \Sigma_1^{-1T} & p_1(2)(x_1 - \mu_2)^T \Sigma_2^{-1T} & p_1(m)(x_1 - \mu_m)^T \Sigma_m^{-1T} \\ p_2(1)(x_2 - \mu_1)^T \Sigma_1^{-1T} & p_2(2)(x_2 - \mu_2)^T \Sigma_2^{-1T} & p_2(m)(x_2 - \mu_m)^T \Sigma_m^{-1T} \\ p_n(1)(x_n - \mu_1)^T \Sigma_1^{-1T} & p_n(2)(x_n - \mu_2)^T \Sigma_2^{-1T} & p_n(m)(x_n - \mu_m)^T \Sigma_m^{-1T} \end{bmatrix}$$

- whose solution is given by the normal equations

$$\begin{bmatrix} \mathbf{P}^T \mathbf{P} & | & \mathbf{P}^T \Delta \\ \hline \hline \Delta^T \mathbf{P} & | & \Delta^T \Delta \end{bmatrix} \cdot \begin{bmatrix} \mathbf{v} \\ \dots \\ \Gamma \end{bmatrix} = \begin{bmatrix} \mathbf{P}^T \mathbf{y} \\ \dots \\ \Delta^T \mathbf{y} \end{bmatrix}$$

- which can be solved through the Cholesky decomposition, since the leftmost matrix is symmetric and positive definite
- Note, however, the computational complexity of the solution
 - Assuming $p = 20$ MFCCs, $m = 128$ Gaussians and $n = 20,000$ frames, computing the block $\Delta^T \Delta$ will require 65×10^9 multiplications, and inversion of the overall leftmost matrix another 3×10^9 multiplications

Discussion

- Methods based on VQ yield impressive results considering their simplicity, but are not robust and the conversions are constrained to be linear combinations of the training codewords
- Methods based on GMMs yield good similarity between converted and target voices, but the synthesis is of low quality (over-smoothing, formant broadening, distortions due to residual and phase)
- Methods based on DFW produce synthesis of good quality, but similarity between converted and target voice is poor because DFW does not change formant amplitudes

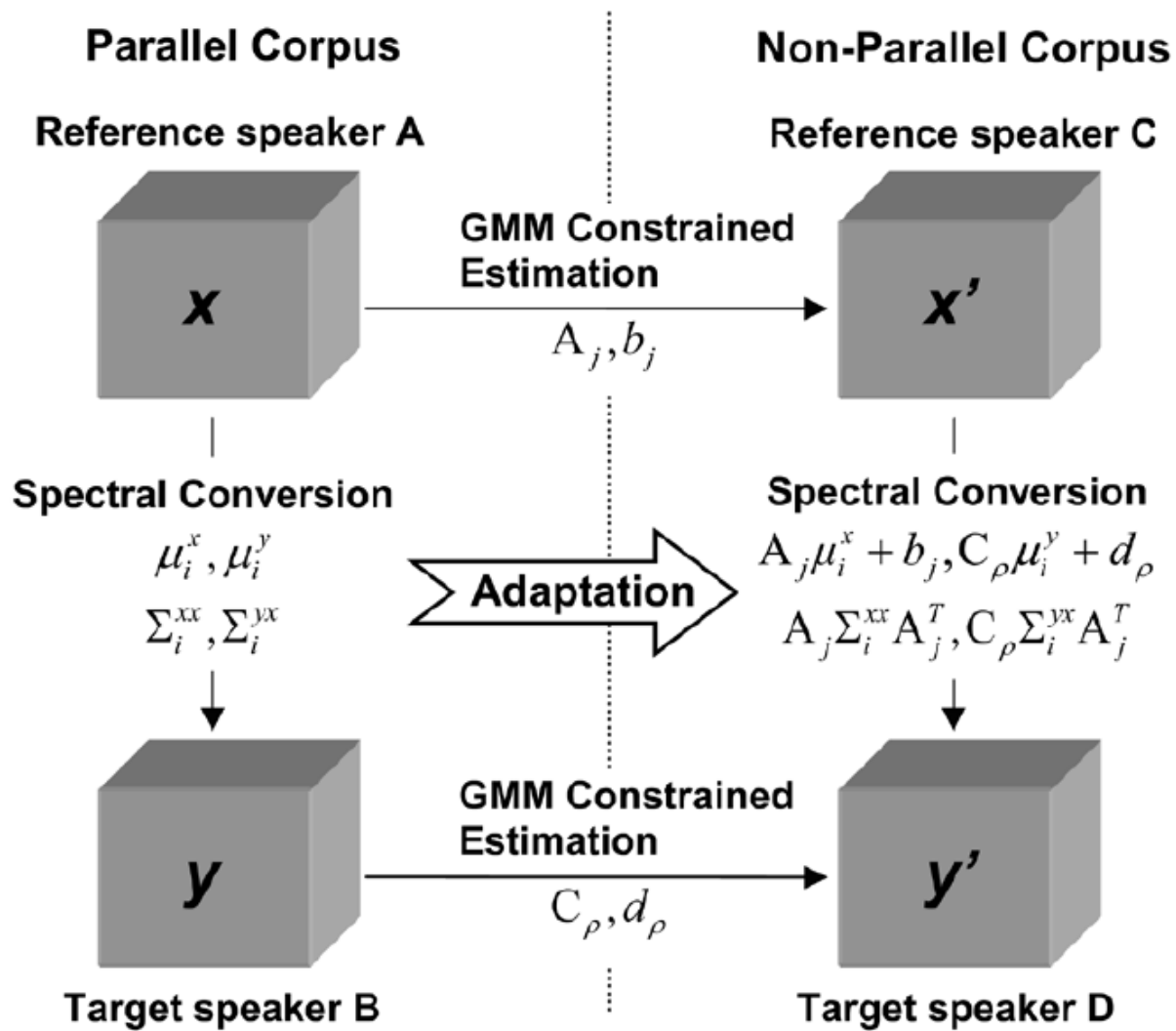
Weighted Frequency Warping (Erro and Moreno, 2007)

- Combines DFW approach of Valbret et al. (1992) with the GMM approach of Stylianou et al. (1998)
- Following Stylianou et al. (1998), build a joint GMM from a corpus of parallel recordings from source and target speaker
- For each GMM cluster, calculate a DFW based on formant positions (formants being estimated from LSFs)
- Once GMM and DFW have been trained, for each test frame x
 - Compute the probability $p_i(x)$ of the frame on each GMM cluster
 - The DFW for frame x is calculated as the weighted average DFW across clusters, with weights equal to $p_i(x)$
- Since DFW modifies the position of formants but not their energy:
 - Correct the energy on different frequency bands (100-300, 300-800, 800-2500, 2500-3500 and 3500-5000Hz) by means of a multiplicative term

Voice conversion w/o parallel corpus

Approach (Mouchtaris et al., 2006)

- Methodology relaxes the constraint of most VC systems, which is that parallel recordings from source and target speakers be available
- Method assumes that only a non-parallel corpus is available for source and target speakers (C and D), but that a parallel corpus is available for a *different* pair of speakers (A and B)
- Build a VC model for speakers A and B using a conventional GMM method
- Build constrained adaptation model from A to C, and from B to D
- Adapt VC model using transformation matrices from the adaptation model (see next page)



[Abe et al., 1988]

Foreign accent conversion

What is a foreign accent?

- Deviations from acoustic (formants) and prosodic (intonation, duration, rate) norms of a language
 - Other cues (e.g. lexical)

Modulation theory [Traunmüller, 1994]

- Speech results from the modulation of a voice quality carrier with linguistic gestures
 - The carrier is defined as the organic aspects of a voice
- To remove a foreign accent, combine the L2 speaker's carrier with the linguistic gestures of a native speaker

How to extract “organic carrier” and “linguistic gestures”?

- Use articulatory information (see Daniel Felps dissertation)
- Acoustics based (vocoding)
 - FD-PSOLA for prosodic modifications
 - SEEVOC and VTLN for segmental modifications

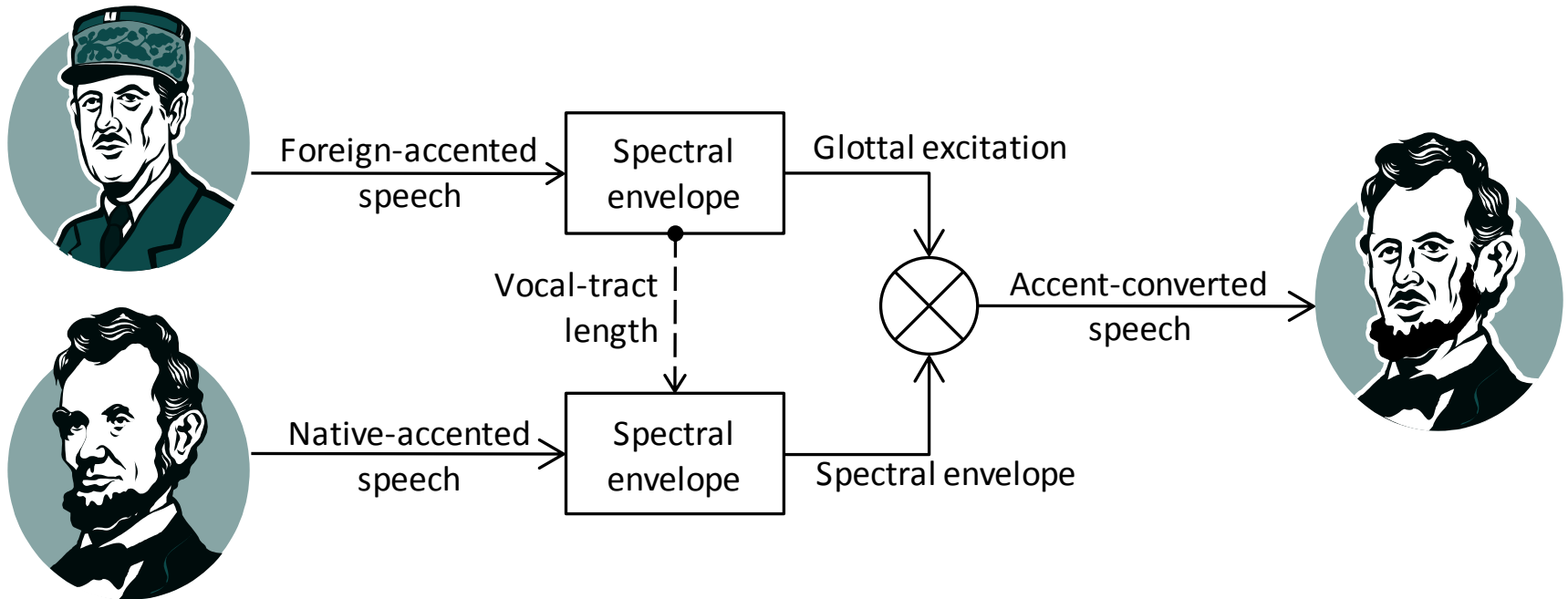
Foreign accent conversion steps

– Prosodic conversion

- Time scaling: done on a phoneme-by-phoneme basis by computing the ratio of source-target durations
- Pitch scaling: done by replacing source $F0$ contour with shifted version of target $F0$ contour

– Segmental conversion

- Combine source glottal excitation with target spectral envelope by means of FD-PSOLA
- To reduce identity cues in target's spectral envelope, we first apply VTLN
- Frequency warp for VTLN is a piecewise linear function defined by the average formant pairs of the two speakers



Perceptual evaluation

- Acoustic quality (MOS scale)
 - Before testing, participants listen to examples with various accepted MOS values
- Foreign accent
 - Rate degree of accented of utterances (7-point EGWA scale: 0:not at all; 2:slightly; 4:quite a bit; 6:extremely)
- Identity ratings
 - Paired test (linguistically different utterances)
 - Are utterances from the same or different speakers? (Y/N)
 - How confident are you? (7-point scale)
 - Unfold responses into a 15-point score

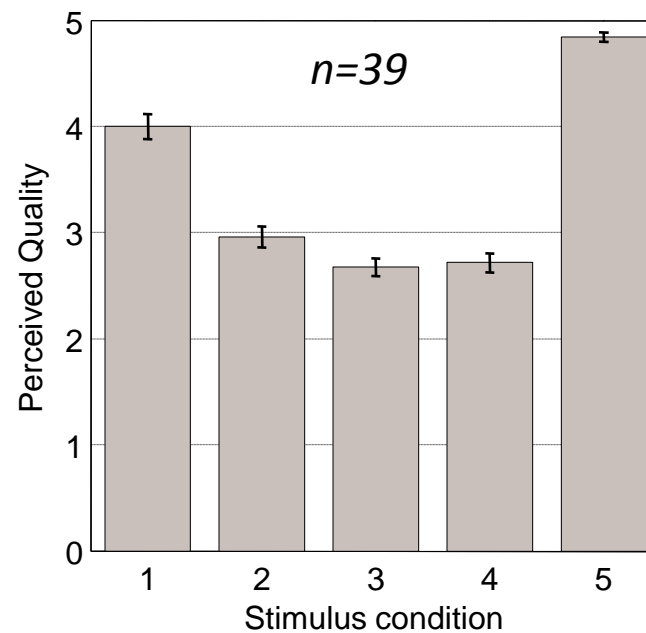
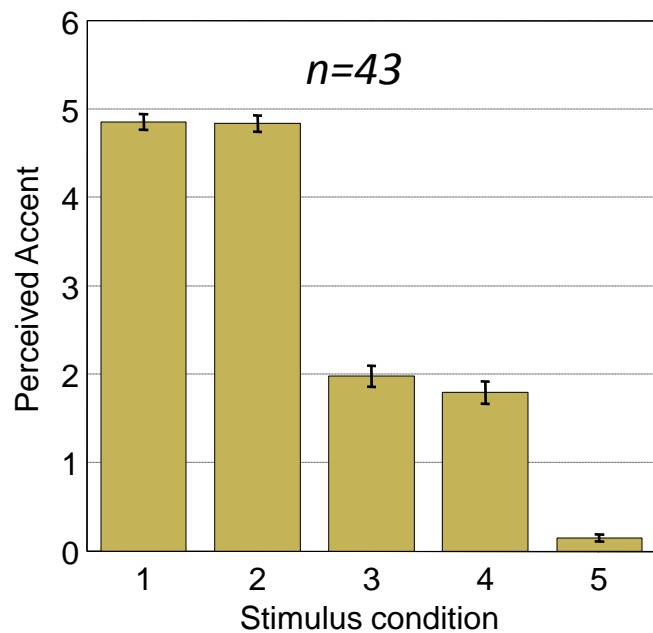
Value	Equivalent meaning
0	Same speaker, very confident
6	Same speaker, not at all confident
7	N/A
8	Different speaker, not at all confident
14	Different speaker, very confident

Experiments

- Subjects
 - 191 undergraduate students in Psychology Dept.
 - US native speakers, no hearing impairments
- Stimuli
 - Two speakers from CMU ARCTIC
 - ksp_indianmale: treated as the foreign speaker
 - rms_usmale2: treated as the native speaker
- Five experimental conditions
 - Learner, prosodic, segmental, p & s, teacher
- The same 20 sentences were chosen for each of five conditions, or 100 unique utterances

Results: perceived accent and quality

#	Stimulus
1	Student utterance
2	Student w/ prosodic conversion
3	Student w/ segmental conversion
4	Student w/ prosodic & segmental conversion
5	Teacher utterance

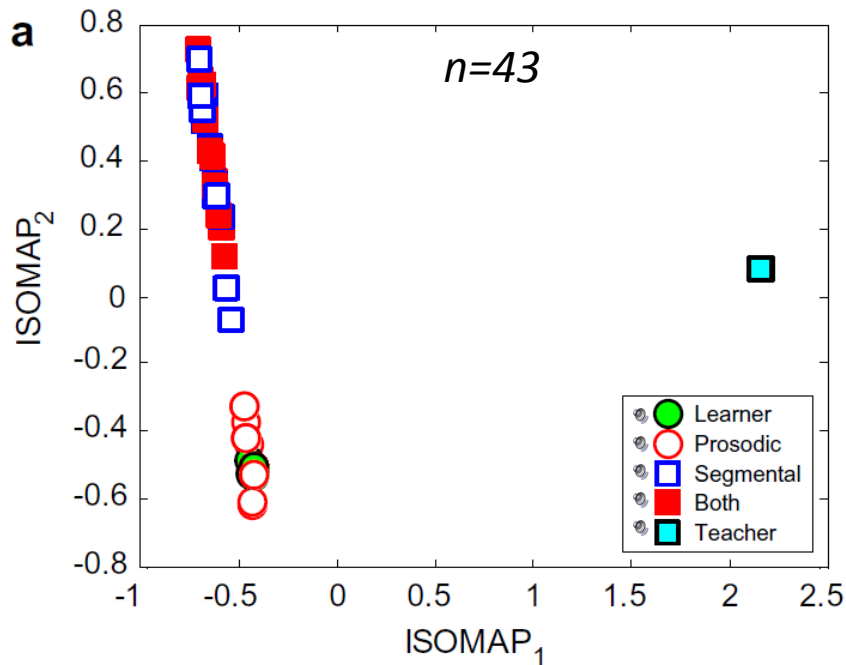


[Felps and Gutierrez-Osuna, 2009]

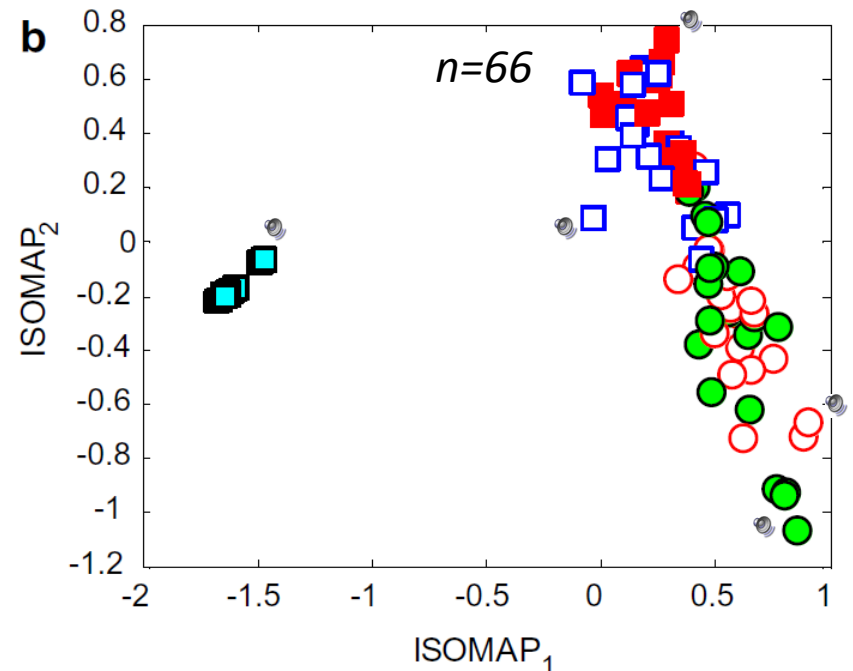
Results: perceived accent and quality

Paired responses mapped into a metric space through multidimensional scaling (ISOMAP)

Normal speech



Reverse speech



[Felps and Gutierrez-Osuna, 2009]