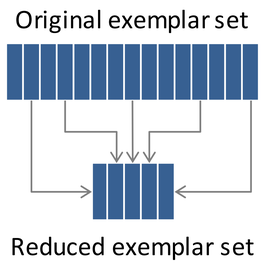
EXEMPLAR SELECTION METHODS IN VOICE CONVERSION

Guanlong Zhao (gzhao@tamu.edu) and Ricardo Gutierrez-Osuna (rgutier@tamu.edu)

lab | COMPUTER SCIENCE & ENGINEERING, TEXAS A&M UNIVERSITY

Abstract

- **Objective:** develop more compact exemplar set for voice conversion (VC)
- **Motivation:**
  - Conventional exemplar-based VC requires large exemplar set to ensure good coverage
  - Factorization is time consuming



- **Solution:** use forward selection and/or backward elimination to reduce the size of an exemplar set

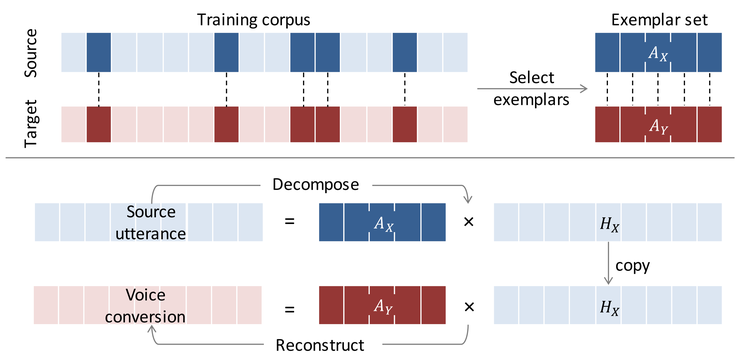
Review: exemplar-based VC methods

- **Conventional NMF:** first, represent utterance X as a weighted sum of exemplars in set A_X

$$X = A_X H_X$$

Then, to generate speech for target speaker Y, combine source weights H_X with target exemplars A_Y

$$\hat{Y} = A_Y H_X$$



- **Joint NMF:** to improve performance [1], one could combine single-frame, high-res (HR) features with multi-frame, low-res (LR) features :

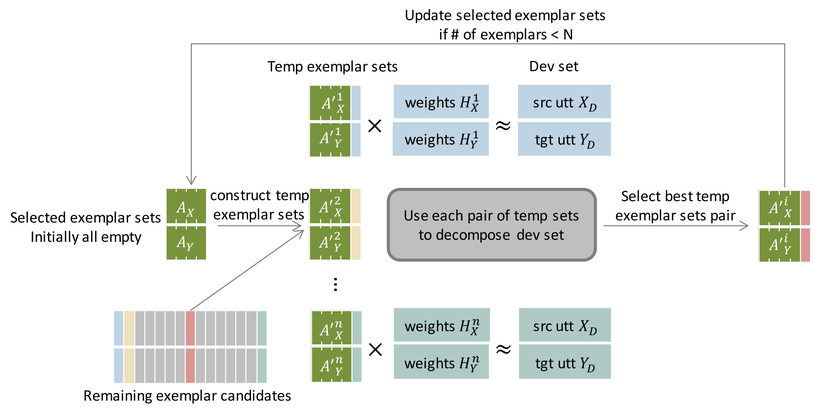
$$\alpha d_{KL}(X^{HR}, A_X^{HR} H) + (1-\alpha) d_{KL}(X^{LR}, A_X^{LR} H) + \lambda \|H\|_1$$

Overview: exemplar selection methods

- Compiling an exemplar set starts by time-aligning source and target utterances in a training corpus
- Conventional approach randomly picks large exemplar sets
- We propose to use sequential selection methods to highlight the most important exemplar candidate pairs

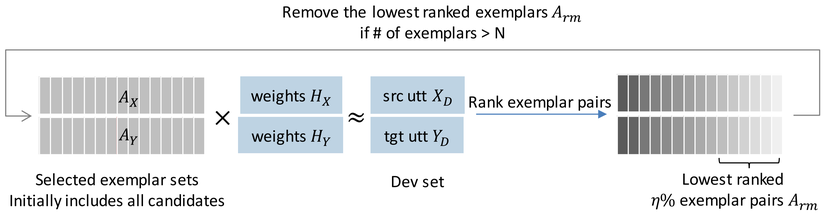
Forward selection (FWD)

Starting with an empty subset, we select the next exemplar as the one that minimizes the joint reconstruction error on source and target utterances from a development set



Backward elimination (BKW)

The algorithm starts with a large subset of exemplars and sequentially removes the exemplars with the lowest activation weights in NMF, computed on a development set



Experimental setup

- **Dataset:** four speakers from ARCTIC: two males (BDL and RMS) and two females (SLT and CLB). For each speaker: 10 utterances for training, 10 for development, and 50 for testing. All sets are phonetically balanced
- **Features:** 513-dim high-res spectra; 23-dim low-res Mel-scale filter bank energies; 25-dim MFCCs
- **VC implementations:** RAND (baseline, uses random exemplars), FWD (uses forward selection to get exemplars), BKW (uses backward elimination to get exemplars)
- **VC tasks:**
  - BDL to RMS (male to male)
  - CLB to SLT (female to female)
  - BDL to SLT (male to female)
  - CLB to RMS (female to male)

Objective evaluation

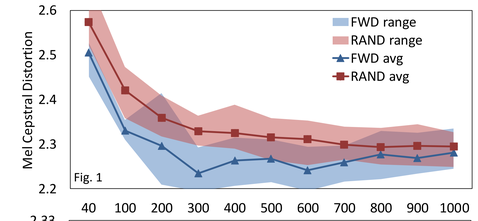


Fig. 1: MCD of FWD vs. the number of exemplars selected. Shaded region indicates min-max range

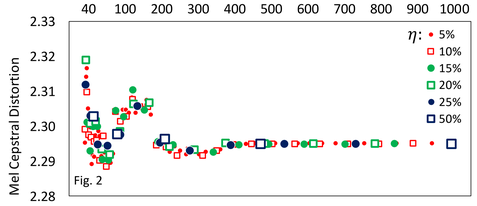


Fig. 2: MCD of BKW vs. the cutoff threshold η. E.g., if $\eta = 5\%$, we remove exemplars that are ranked among the lowest 5% in each iteration

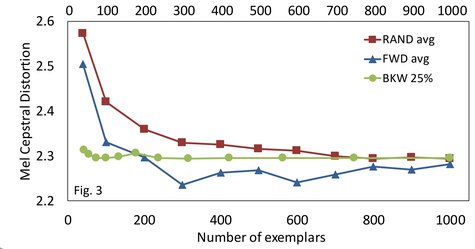


Fig. 3: Comparison against random selection. RAND was averaged over 8 runs, BKW used $\eta = 25\%$

Subjective evaluation

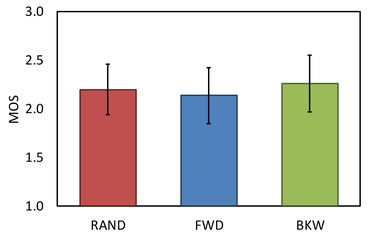


Fig. 4: Mean opinion score (MOS) measures acoustic quality. We used 1000, 200, and 204 exemplars for RAND, FWD, and BKW, respectively

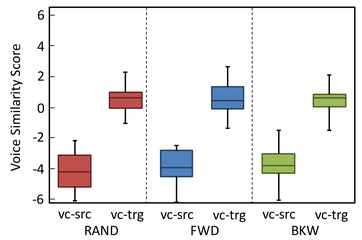


Fig. 5: Voice similarity score [2] measures speaker identity. Same-gender VC VSS=5.3 ("same speaker"); cross-gender VC VSS=−4.5 ("different speakers")

Discussion

- FWD builds a compact exemplar set by reducing the joint reconstruction error of source and target utterances
- BKW excludes exemplars that contribute the least to the factorization
- Both strategies can significantly reduce the number of exemplars, without sacrificing VC performance
- **Future work:** combine both strategies; improve the computational efficiency of FWD

References

[1] Z. Wu, et al., "Exemplar-based voice conversion using joint nonnegative matrix factorization," *Multimedia Tools and Applications*, vol. 74, pp. 9943-9958, 2015.

[2] D. Felps, et al., "Developing objective measures of foreign-accent conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1030-1040, 2010.