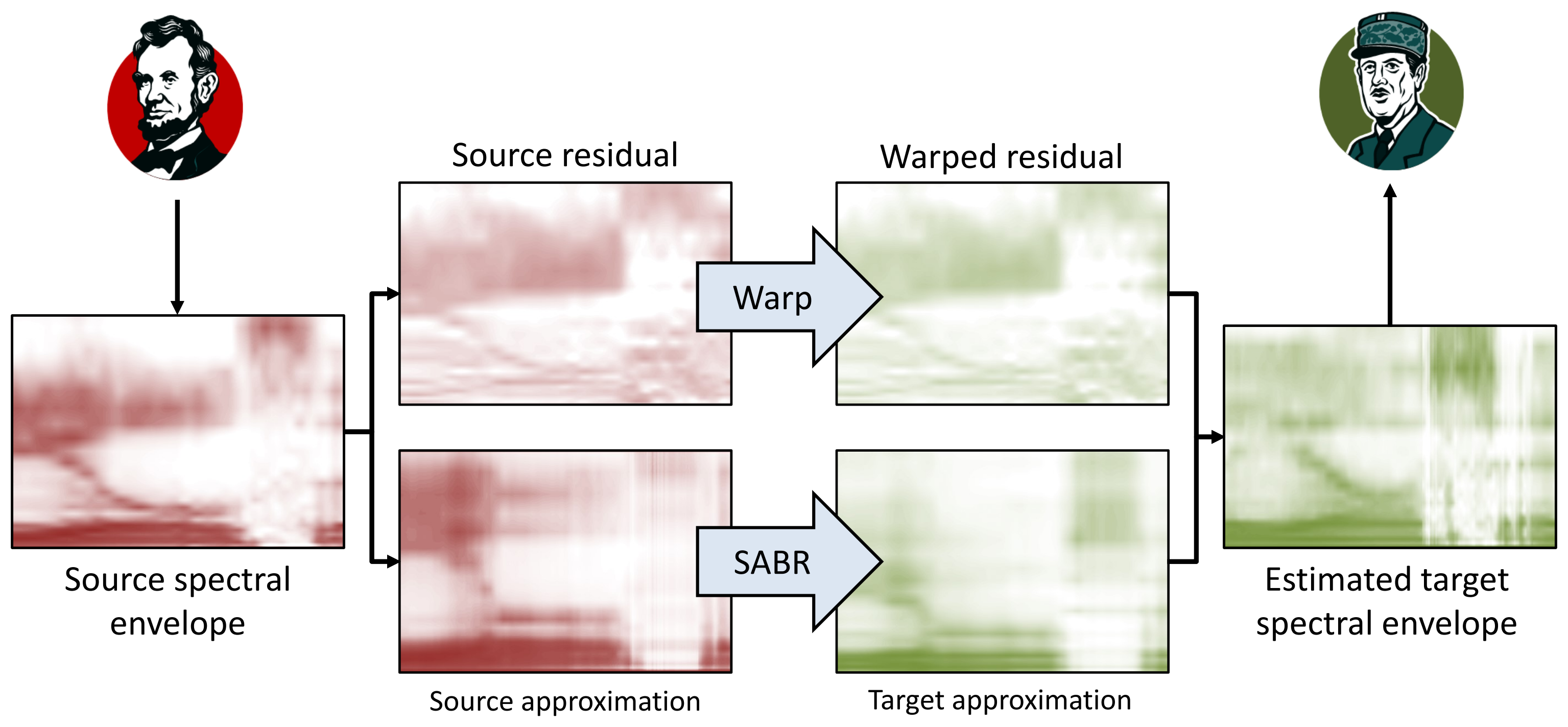
Voice Conversion through Residual Warping in a Sparse, Anchor-Based Representation of Speech

Christopher Liberatore, Guanlong Zhao, and Ricardo Gutierrez-Osuna

COMPUTER SCIENCE & ENGINEERING
TEXAS A&M UNIVERSITY

Abstract

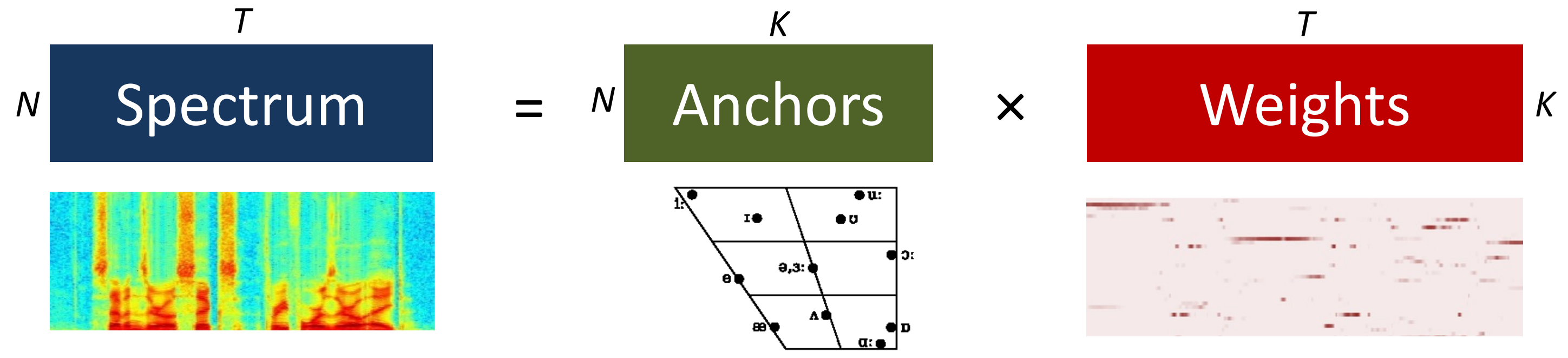
- **Objective:** Improve the synthesis quality of voice conversion in a sparse, anchor-based representation of speech (SABR) [1]
- **Motivation:** Just using anchors in voice conversion results in low-quality synthesis due to a large source residual that was not used in voice conversion
- **Problem:** The source residual needs to be considered, but may contain speaker identity and needs to be converted to the target speaker
- **Solution:** Use source and target anchors to learn warping functions to warp the source residual to the target



Sparse Anchor-Based Representation (SABR)

SABR represents an utterance X using phonetic anchors A and a weight vector W. A residual term R accounts for the error in the SABR approximation:

$$X = AW + R \quad \text{(eq. 1)}$$



We compute W using Lasso regression:

$$\min_{W} \|X - A_S W\|^2 + \lambda \|W\|_1 \; s.t. W \in [0,1] \quad \text{(eq. 2)}$$

We use W to estimate the target envelope using target anchors A_T:

$$X_T \cong A_T W \quad \text{(eq. 3)}$$

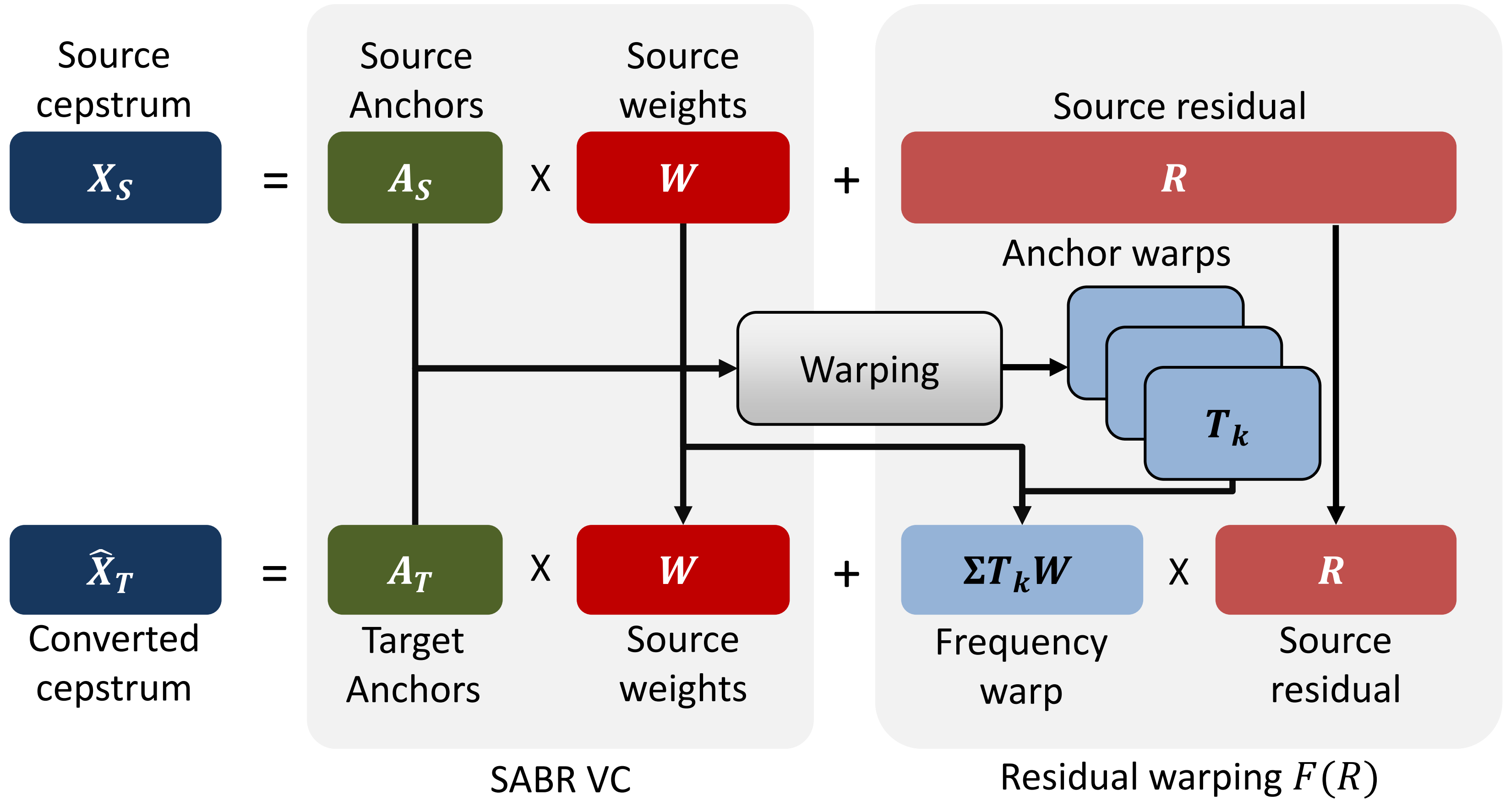
This approximation needs to account for the residual. We cannot just add R from eq. 1. Therefore, we use the anchors to learn a warping function:

$$X_T \cong A_T W + F_R(R) \quad \text{(eq. 4)}$$

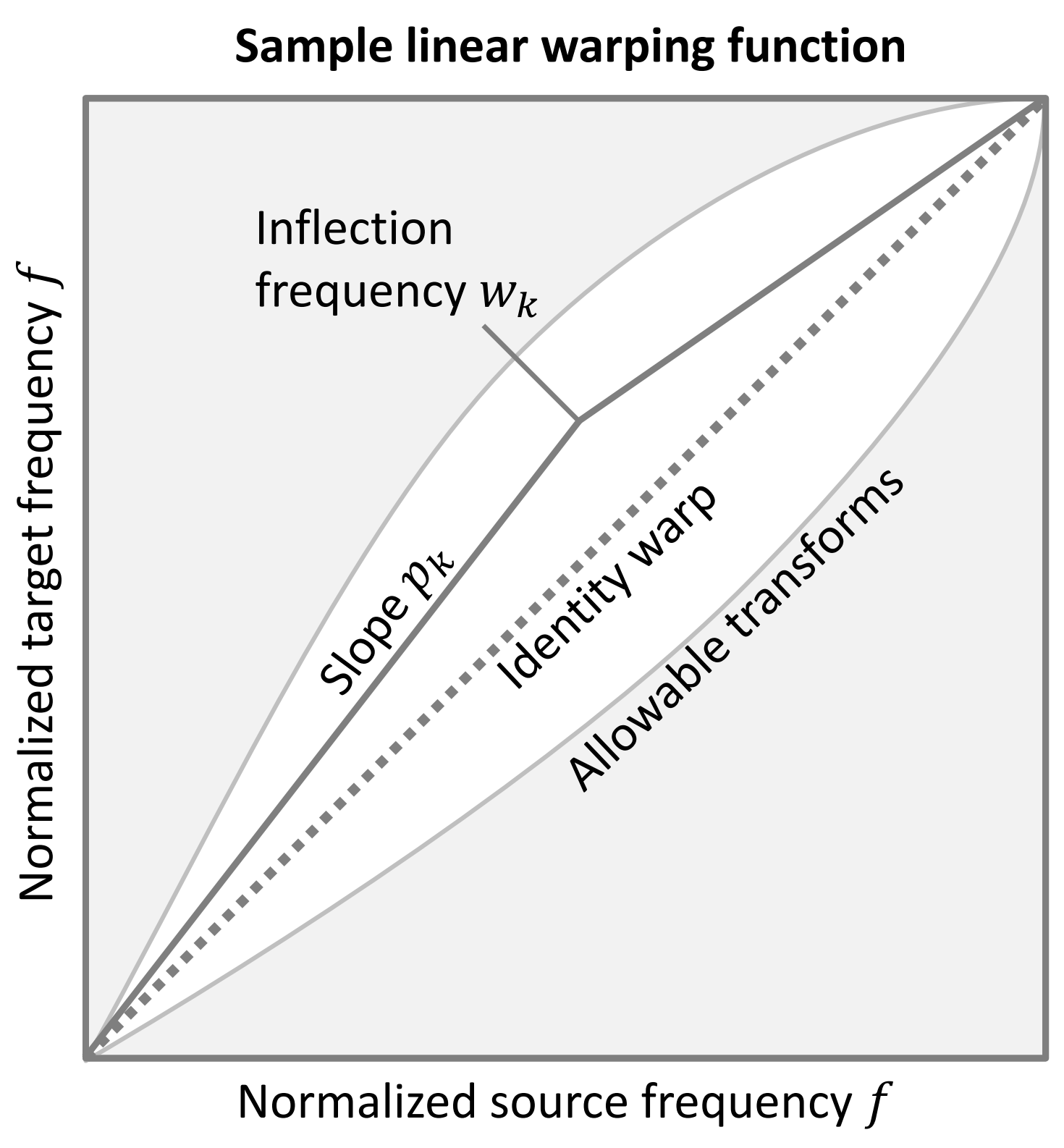
Residual Warping

- Using the source and target anchors, we estimate a transform T_k for each anchor k minimizing $\left|\left|A_S^k - T_k A_T^k\right|\right|_2^2$
- **Vocal Tract Length Normalization (VTLN) functions** can be represented as linear transforms of MFCCs [2]
- For VTLN, we use a **piecewise linear warping function** with parameters ω_k and p_k (inflection frequency and slope)
- For each frame, we use the weights to learn a warping function from the learned anchor warps
- Using the weights W and transforms T, the full voice conversion method including residual warping becomes

$$\hat{X}_i = A_T W_i + \left(\sum_{\forall k} W_{i,k} T_k(\omega_k, p_k) \right) R_i \quad \text{(eq. 5)}$$



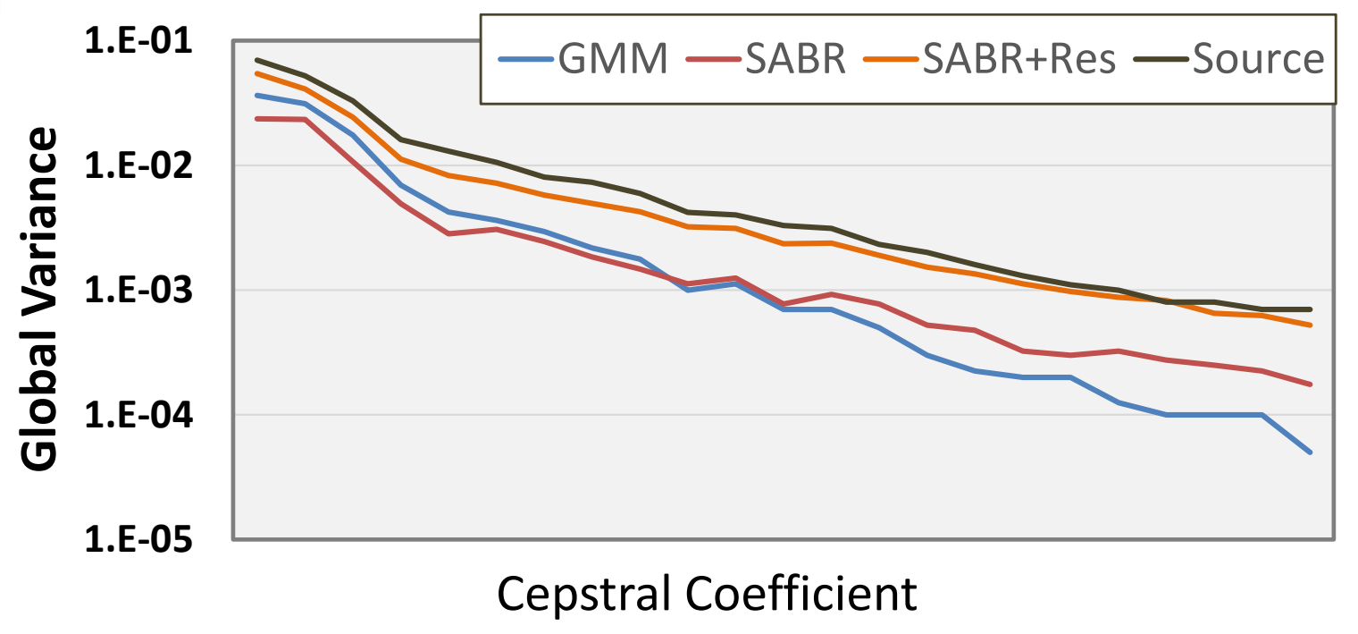
Corpus and Parameters


Sample linear warping function
Inflection frequency w_k
Slope p_k
Identity warp
Allowable transforms
Normalized target frequency f
Normalized source frequency f

- **Corpus:** ARCTIC speech corpus, speakers BDL, CLB, RMS, and SLT
- **Anchor selection:** one anchor for each phoneme by computing the centroid of MFCCs from frames with that label
- **Anchor features:** $MFCC_{1-24}$ (excluding energy)
- **Sparsity penalty:** $\lambda = 0.025$
- **Warp parameters:** Inflection frequency $w_k \in [0.4\ 0.8]$ Slope parameter $p_k \in [0.8, 1.2]$

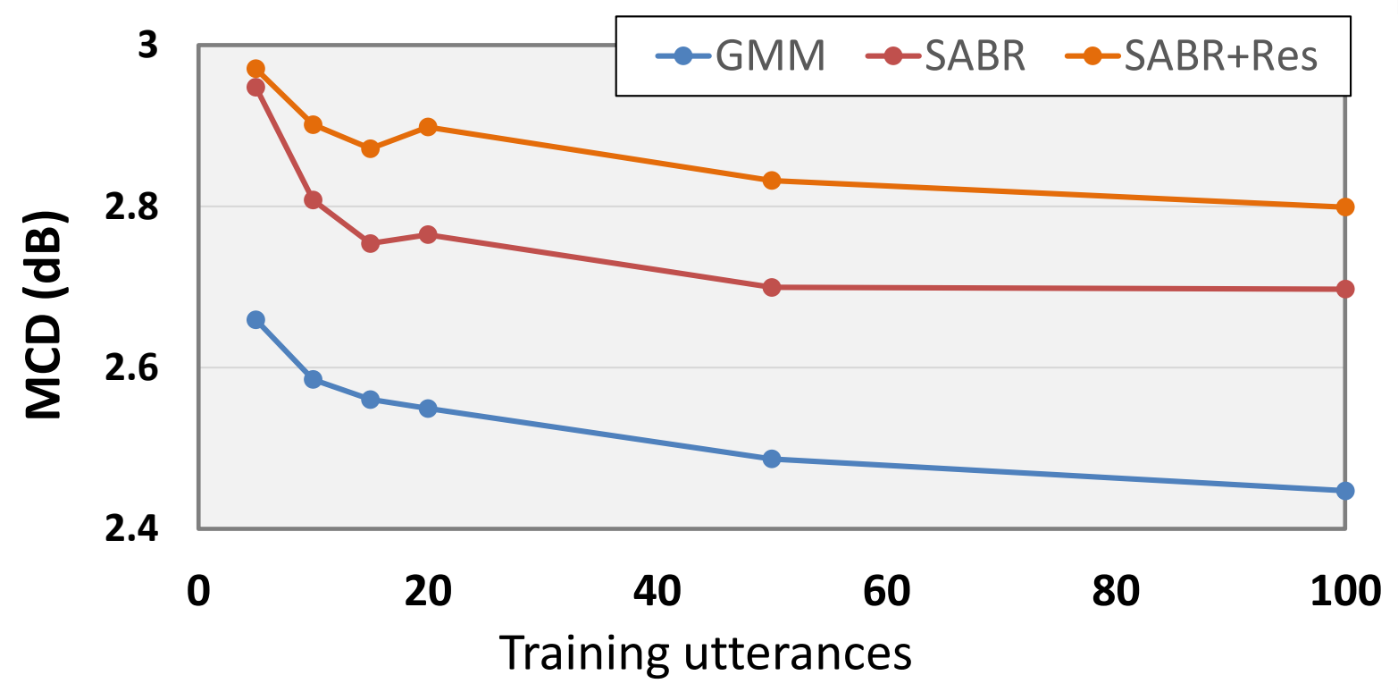
Experiments

We compared three different systems: **SABR** (eq. 3), **SABR+Res** (eq. 5), and a 40-mixture **GMM** with diagonal covariances. Models were trained on the same utterances.



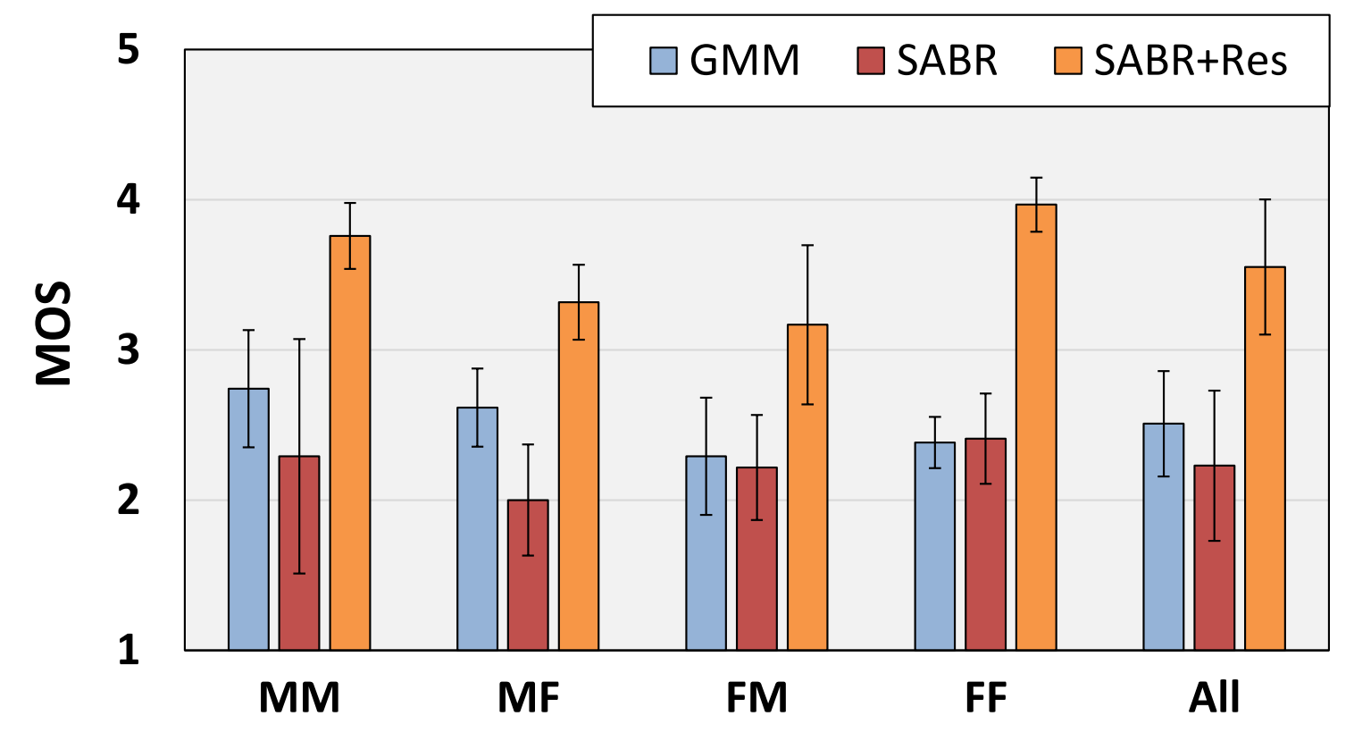
Cepstral variance

- Global variance of MFCCs as measure of quality
- **SABR+Res** approached source variance



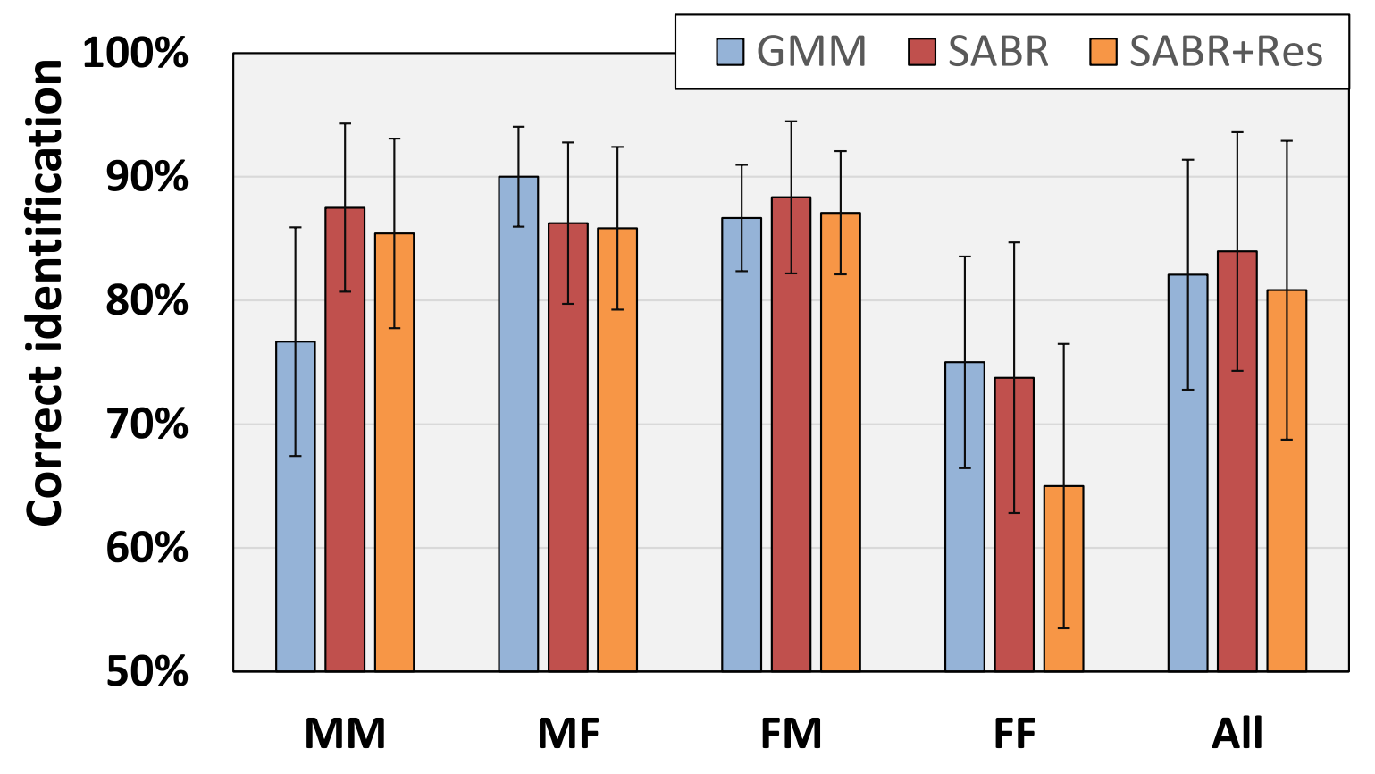
Objective VC

- Compared against number of training utterances
- Residual adds ~0.1 dB error to est. target spectrum



Mean Opinion Score

- 15 training utterances
- Warped residual *significantly* increased MOS (SABR: 2.2; GMM: 2.5; SABR+Res: 3.6; $p > 0.01$)



XAB Identity Test

- **SABR+Res** performed at least as well as **GMM** ($p = 0.35$)
- F-F performance low; CLB/SLT are very similar

Conclusions

- **Residual warping** improved rated acoustic quality though VC error increased
- The increased VC error **did not affect** the ability for a listener to perceive the identity of the speaker

Future work:

1. Determine ideal anchor sets, as some phoneme classes may be ill-suited for single-vector anchors (e.g. stops).
2. Add temporal smoothness constraints in the objective function via the Fused Lasso [3]

References

[1] C. Liberatore, S. Aryal, *et al.* “SABR: Sparse, Anchor-Based Representation of the Speech Signal,” *Interspeech* 2015.

[2] S. Panchapagesan and A. Alwan, "Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC," *Computer Speech & Language,* vol. 23, no. 1, pp. 42-64, 2009.

[3] Tibshirani, Ryan J., and Jonathan Taylor. "The Solution Path of the Generalized Lasso." The Annals of Statistics (2011): 1335-1371.

Scan QR code for audio samples