VOICE CONVERSION THROUGH RESIDUAL WARPING IN A SPARSE, ANCHOR-BASED REPRESENTATION OF SPEECH

Christopher Liberatore, Guanlong Zhao, Ricardo Gutierrez-Osuna

Texas A&M University, College Station, Texas, USA

{cliberatore, gzhao, rgutier}@tamu.edu

ABSTRACT

In previous work we presented a Sparse, Anchor-Based Representation of speech (SABR) that uses phonemic "anchors" to represent an utterance with a set of sparse non-negative weights. SABR is speaker-independent: combining weights from a source speaker with anchors from a target speaker can be used for voice conversion. Here, we present an extension of the original SABR that significantly improves voice conversion synthesis. Namely, we take the residual signal from the SABR decomposition of the source speaker's utterance, and warp it to the target speaker's space using a weighted warping function learned from pairs of source-target anchors. Using subjective and objective evaluations, we examine the performance of adding the warped residual (SABR+Res) to the original synthesis (SABR). Specifically, listeners rated SABR+Res with an average mean opinion score (MOS) of 3.6, a significant improvement compared to 2.2 MOS for SABR alone ($p < 0.01$) and 2.5 MOS for a baseline GMM method ($p < 0.01$). In an XAB speaker identity test, listeners correctly identified the identity of SABR+Res (81%) and SABR (84%) as frequently as a GMM method (82%) ($p = 0.70$, $p = 0.35$). These results indicate that adding the warped residual can dramatically improve synthesis while retaining the desirable independent qualities of SABR models.

Index Terms**--**sparse coding, voice conversion, residual, dynamic frequency warping, weighted frequency warping

1. INTRODUCTION

Voice conversion (VC) is the process of taking an utterance from one speaker and converting it to sound as if another speaker had produced it, i.e., VC combines the linguistic content of the source speaker's utterance with the voice quality of the target speaker. VC can be useful in a variety of contexts, from changing the identity of speakers in text-to-speech systems [1] to generating "golden speakers" in pronunciation training [2]. VC often requires large, parallel corpora [3], though some methods [4] relax these assumptions.

In prior work [5], we presented one such method. Termed SABR (Sparse, Anchor-Based Representation of speech), this method represents an utterance as a sparse, nonnegative linear combination of phoneme "anchors"—each anchor being the acoustic centroid for a phoneme class. SABR has several desirable properties: it requires a very small training corpus, it does not require parallel recordings, and it does not require training for each pair of source-target speakers. However, the compact anchor set lacks the variability to represent the details of an utterance. As a result, the utterance representation has a muffled quality.

This paper proposes a method that significantly improves the VC synthesis quality of SABR by warping the residual signal of the source utterance to match the acoustic space of the target speaker. The method operates as follows. First, we compute a piecewise linear warping function for each pair of source-target anchors, i.e., one function per phoneme; this step needs to be performed only once, during the initial training phase. To convert a new source utterance, we use Lasso [6] to compute the SABR weights (relative to the source's anchors) and the corresponding residual (i.e., the reconstruction error). Next, for each frame in the source utterance, we compute a weighted warping function as the sum of each anchor's warping function multiplied by its weight. In the final step, we estimate the spectrum of a target utterance by multiplying the source weights by the target anchors, and add the warped source residual using the warping function learned previously. We evaluate the performance of the method using objective (e.g. Mel cepstral distortion (MCD) and cepstral variance) and subjective (mean opinion scores, XAB preference tests) measures. Our results indicate that the addition of the warped residual greatly enhances the audio quality while still generating synthesized acoustics with the voice quality of the target speaker.

The rest of this paper is organized as follows. First, we review prior VC methods that are most closely related to our approach. Then, we briefly describe the original SABR algorithm and derive the proposed residual-warping method. Next, we present results on subjective and objective experiments, using utterances from the ARCTIC database [7]. Finally, we discuss the implications of this method and provide directions for future work.

2. PRIOR WORK

A common way to perform VC is through statistical learning, most commonly Gaussian Mixture Models [3, 8]. These methods perform regression to map acoustic features (e.g., MFCCs) from source to target speaker, typically using parallel, time-aligned data. However, these methods can suffer from spectral over-smoothing and onerous data collection requirements.

To improve the quality of GMM-based methods, Erro *et al.* [9] proposed frequency warping and amplitude scaling (FW+AS). For each mixture, a warping function and amplitude scaling vector was learned to map from the source to the target speaker. During conversion, instead of using the conditional probability of the GMM to estimate the target spectral envelope, the conditional probability was used to estimate a warping function and amplitude scaling of the source utterance. Their method outperformed a baseline GMM in terms of decreased spectral distortion and higher preference ratings from listeners. Godoy *et al.* [1] presented a similar method, but removed the requirement for parallel utterances, instead building a phonemic GMM. For each phoneme class, the authors computed ideal frequency warping and amplitude scaling functions; however, the amplitude scaling was estimated from the residual of the warped source and target acoustics. The authors found that listeners preferred their method to standard GMM regression, even though it led to higher spectral distortion than a traditional GMM-regression method.

An alternative to GMMs has emerged in recent years [4, 10]. Known as exemplar-based voice conversion, these methods use nonnegative matrix factorization (NMF) to represent utterances as a linear combination of exemplars (i.e., short-time spectra) with an activation matrix. These methods require an initial pairing of source exemplars with target exemplars. Afterwards, for each new source utterance, these methods use NMF to compute an activation matrix (relative to the source exemplars) and then combine it with the corresponding target exemplars. Listening tests by Wu et al. [4] and Aihara et al. [10] have shown a preference for exemplar-based methods over comparable GMMs.

In subsequent work, Wu *et al.* [11] incorporated frequency warping into exemplar-based voice conversion. First, pairs of source-target exemplars are used to train warping functions and residual exemplars. Then, the activation matrix is used to compute a warping function that warps the source utterance to the target speaker, retaining the original spectral detail; a residual spectrogram computed from the activation matrix and residual exemplars, and then added to the warped source Listeners preferred the exemplar-based warping method to a GMM-based warping method.

Our proposed work differs from these prior methods in several respects. First, our exemplar set is significantly smaller since it is derived (one-to-one) from phonemic labels; because we use labels, the parallel constraint is also removed. Additionally, we do not compute residual exemplars, instead electing to warp the residual as opposed to the source utterance. Finally, since SABR anchors are tied to specific phonemes, the resulting SABR weights are interpretable.

3. METHODS

3.1. Sparse Anchor Based Representation (SABR)

SABR represents an utterance as a sparse weighted sum of speaker-dependent phonemic anchors [5]. These anchors are obtained in a semi-supervised manner (e.g., force-alignment) or through manual annotation. This modeling allows us to learn a speaker-independent representation for VC with minimal training data. For each speaker, we learn a phoneme anchor A^k by choosing the centroid frame from all training samples with phoneme label k. For a given source utterance with N acoustic features (e.g., MFCCs) and T frames, $X \in \mathcal{R}^{N\times T}$ and a source anchor set $A_S \in \mathcal{R}^{N\times K}$ of K phonemes, SABR approximates the utterance as a weighted sum of the anchors $W \in \mathcal{R}^{K\times T}$:

$$X \cong A_S W \tag{1}$$

To solve for W, SABR uses the Least Angle Regression algorithm[12] to solve the LASSO [6] :

$$W = argmin \left|\left|X - AW\right|\right|^2 s.t. |W|_1 \leq 1, W \geq 0 \tag{2}$$

The weight matrix W can then be used to estimate a target speaker's spectrum as:

$$\hat{Y} = A_T W \tag{3}$$

where $A_T \in \mathcal{R}^{NxK}$ is a set of anchors from the target speaker.

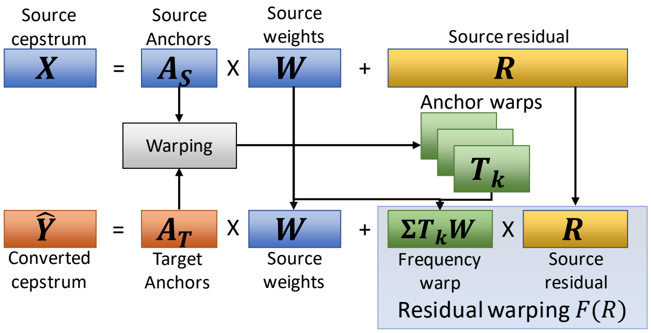

Figure 1: Overview of the proposed residual warping method.

3.2. Residual warping in SABR voice conversion

Eq. (1) can be viewed as an approximation to the original spectrum X that only captures phonological information, and ignores phonetic information in the utterance:

$$X = A_S W + R \tag{4}$$

where R is the residual signal, which contains spectral detail that is specific to the source speaker. As such, this residual cannot be added to eq. (3) without first transforming it into the target speaker's space:

$$\hat{Y} = A_T W + F(R) \tag{5}$$

where $F(R)$ is a residual mapping function. Helander et al. [13] proposed a kernel-based Partial Least Squares mapping of the residual, but they required parallel data. To remove this requirement, we instead use the weight matrix W to estimate a mapping function for the residual. Our overall approach for mapping the two residuals is illustrated in Figure 1.

Following Panchapagesan and Alwan [14], we use a piecewise linear warping function with two free parameters: an inflection point ω_0 (normalized frequency), and a slope parameter p, which is the slope of the warping from 0 to ω_0:

$$f_{pw}(\omega; \omega_0, p) = \begin{cases} p\omega, & 0 \leq \omega \leq \omega_0 \\ p\omega_0 + \left(\frac{1 - p\omega_0}{1 - \omega_0}\right)(\omega - \omega_0), & \omega_0 < \omega \leq 1 \end{cases} \tag{6}$$

When using cepstral coefficients, the transform in eq. (6) can be expressed as a linear transform. Following [14], we compute this transform as a product of a Discrete Cosine Transform (DCT) matrix C and its warped inverse (IDCT) $\hat{C}$. Assuming M filters in an MFCC filterbank, N cepstral coefficients, and a warping function $f(\omega)$, matrices $C \in \mathcal{R}^{N\times M}$ and $\hat{C} \in \mathcal{R}^{M\times N}$ can be computed as:

$$C_{m,k}{}^T = [\alpha_k \cos(\pi k \omega_m)]_{\substack{1\leq m\leq M \\ 0\leq k\leq N-1}} \tag{7}$$

$$\hat{C}_{m,k} = \left[\alpha_k \cos\left(\pi k f(\omega_m)\right)\right]_{\substack{1\leq m\leq M \\ 0\leq k\leq N-1}} \tag{8}$$

where α_k is a term to ensure that the DCT is unitary, and ω_m is the normalized frequency for the mth Mel filter. The linear warping of the MFCCs is $T = C\hat{C}$, where $T \in \mathcal{R}^{N\times N}$. Substituting $f_{pw}(\cdot)$ from eq. (6) into eq. (8), the transform becomes a function of ω_0 and p:

$$T(\omega_0, p) = C\hat{C}(\omega_0, p) \tag{9}$$

For each pair of source-target anchors A_S^k and A_T^k, we create a transform T_k by selecting ω_0 and p to minimize the SSE of the transformed source and target anchors:

$$T_k = \underset{T(\omega_0,p)}{\operatorname{argmin}} \sum \left(T(\omega_0,p)A_S^k - A_T^k\right)^2 \tag{10}$$

Following [15], we constrain the inflection frequency $\omega_0 \in [0.4,0.8]$ and the warping slope $p \in [0.8,1.2]$. The resulting residual warping VC method is similar to Weighted Frequency Warping [16].

The final transform of the residual (i.e., $F(R)$ in eq. (5)) is the weighted sum of the individual anchor transforms T_k. We add a single row $W_{k+1} = 1 - \left|\left|W_{1\ldots k}\right|\right|_1$ to ensure the weights sum to 1 and set the corresponding warp $T_{k+1} = I$. For each source frame X_i, SABR weight vector W_i, and the frame residual R_i, we estimate the target speaker's spectrum $\hat{Y}_i$ as:

$$\hat{Y}_i = A_T W_i + \left(\sum_{k=1}^{K+1} W_{i,k} T_k\right) R_i \tag{11}$$

Because of the sparsity imposed in eq. (2), the resulting residual transform matrix favors weights on or near the diagonal, a cepstral VTLN property noted by Pitz and Ney [15].

4. EXPERIMENTS

4.1. Corpus

For our experiments, we used data from the ARCTIC speech corpus [7] that includes phonetic transcriptions for each utterance. We used the four American English speakers in ARCTIC: BDL, CLB, RMS, and SLT. We used STRAIGHT [17] with default settings to extract aperiodicity, fundamental frequency and spectral envelope, then computed a 24-dimension MFCC vector (25 filterbanks, 24 coefficients not including $MFCC_0$ (energy), 8 KHz cutoff) from the spectral envelope. We assign each acoustic frame a phonetic label based on the ARCTIC transcription.

4.2. Voice conversion model design

We evaluate the proposed method (SABR+Res) against two baseline systems: the original SABR method without residual compensation [5], and a baseline GMM conversion system [3]. We elected to not perform a more complex form of GMM-based VC (such as adding MLPG [8]) as these methods would not consistently converge with limited training data. To show that our approach does not require parallel training data, we built the source and target SABR models on mutually-exclusive sets of utterances from ARCTIC's "A" set. Utterances were chosen in such a way as to maximize phoneme variability. GMM models were trained using the same utterances used to train the source SABR model and the time-aligned parallel utterances of the target; thus, the GMM models had a slight advantage compared to SABR and SABR+Res. GMMs were set to 40 mixtures to have comparable complexity with SABR, and diagonal covariances. Following prior work [5], we perform log-mean and variance scaling of the source F0 contour to match the range of the target F0.

We examined a subset of speaker pairs—one for each gender conversion direction: M-M (BDL to RMS), F-F (SLT to CLB), M-F (BDL to CLB) and F-M (SLT to BDL); Following [5], for perceptual experiments we recruited listeners through Amazon's online crowdsourcing tool Mechanical Turk. For comparability, we only perform objective evaluations on the same four speaker pairs. In all instances, we perform voice conversion on utterances in the "B" set of ARCTIC.

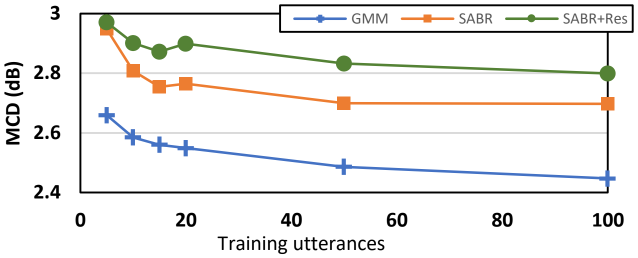

Figure 2: Average Mel-cepstral distortion of SABR, SABR+Res, and the difference between the source and target speakers. The slight increase in SABR MCD between 15 and 20 utterances is due to sensitivity in training.

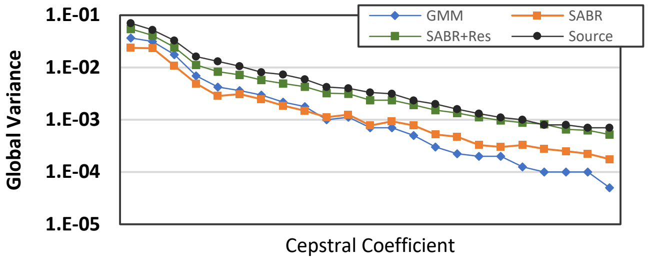

Figure 3: Average Mel-cepstral variance of SABR, SABR+Res, and original utterances. The SABR+Res VC utterances have global variances approaching that of the source utterances.

4.3. Objective experiments

4.3.1. Mel-Cepstral Distortion (MCD)

In a first experiment, we measured the MCD of voice-converted utterances with that of time-aligned target data. For each pair of speakers and VC method, we trained the models using different amounts of training data, ranging from 5 to 100 utterances. Results are shown in Figure 2. A slight uptick in the MCD can be seen in the SABR models as they transition from 15 to 20 utterances. This is due to the fact that SABR anchors are computed as centroids, and given small amounts of training data (roughly 40 seconds) the anchors can shift significantly. GMM models have lower MCD likely because they are trained on time-aligned source-target data, and the training procedure fits to the distribution of the data, not phoneme labels as in SABR. For the next experiments, we used the models trained using 10 utterances, as that was the number of utterances needed to consistently have intelligible audio quality.

4.3.2. Mel-Cepstral Variance

Following prior studies [8, 9], we examined the global variance of the MFCCs as a measure of acoustic quality, as MCD alone may not fully characterize the conversion quality. Figure 3 shows the global variance of each cepstral coefficient for the VC methods and the original ARCTIC utterances. SABR+Res approaches the global variance of the original utterances, besting SABR and GMM. The higher variance of each cepstral coefficient is indicative of better acoustic quality.

4.4. Subjective evaluation

4.4.1. Mean Opinion Score

To evaluate acoustic quality, we used the standard 5-point mean opinion score (MOS; 1-bad, 5-excellent). We recruited 25 native English speakers and asked them to rate 60 utterances: 5 utterances for each VC pair and conversion method. We used 8 unmodified utterances to detect if participants were cheating, removing them if so [18]. Utterances were randomly ordered. Results are shown in Figure 4. Listeners consistently rated SABR+Res as being superior to either SABR or GMM ($p < 0.01$ in both cases).

4.4.2. XAB Preference Test

In a final experiment, we used an XAB test to measure the speaker individuality. Namely, we recruited 24 participants to listen to an utterance X generated using VC as well as two other utterances A and B from the source and target speakers, and then asked if A or B was closer to X in terms of speaker identity. For each VC pair and conversion method, we generated 10 utterances and paired them with different utterances from the source and target speakers. Results are shown in Figure 5. Listeners rated SABR and SABR+Res the same accuracy as the GMM method ($p = 0.70, p = 0.35$); same-gender conversion was slightly reduced in some cases.

5. DISCUSSION

Our experiments show that SABR+Res can dramatically improve VC quality compared to SABR synthesis. Adding the warped residual improves MOS significantly (from 2.2 to 3.5, $p < 0.01$), in agreement with prior studies that use residuals and frequency warping [1, 9, 19]. Moreover, adding the warped residual increases the cepstral variance of the synthesized utterances, bringing it close to that of the original utterances (see Figure 3). Though the MCD increases when the warped residual is added to the SABR voice conversion, we note that the increased distortion (0.12 dB) is smaller than the average magnitude of the residual (1.7 dB), suggesting that residual warping adds a significant amount of "correct" detail. The MCD of GMM VC is lower than SABR likely because of the GMM is fitted to time-aligned source-target data, something SABR does not use. Samples of SABR and SABR+Res spectral envelopes compared with target speech can be seen in Figure 6.

On average, adding the warped residual to the original SABR utterance did not affect the ability for listeners to correctly identify the target speaker, compared with SABR ($p = 0.35$) or GMM-based methods ($p = 0.70$). Identification rates compared favorably with prior warping literature ([1, 9, 11, 16]). Same-gender SABR+Res conversions had lower performance than cross-gender conversions, and we suggest two explanations: first, pitch range is a strong cue to speaker identity in cross-gender conversion, so listeners may focus on pitch differences during identification, making the cross-gender task easier. Second, the warped residual may still retain some of the source speaker identity. This explanation shows predominantly in the F-F conversion, as the two female speakers had very similar voices prior to conversion. If the residual still retained some of the source identity, it could cause confusion between two already-similar speakers[1].

6. CONCLUSION AND FUTURE WORK

We presented a modification to SABR that significantly improves synthesis quality for VC in scenarios where training data is limited. This improvement requires no additional parameters to the original SABR model [5], so the approach remains highly interpretable. Using frequency warping functions learned from source and target anchors, we warped the source residual to the target speaker's space, and added it to the estimated target speaker's spectrum. Through subjective and objective experiments, we found the overall synthesis quality improved dramatically while retaining the ability to capture the voice quality of the target speaker. The resulting synthesis also compares favorably with a GMM-based conversion method, but does not require parallel data.

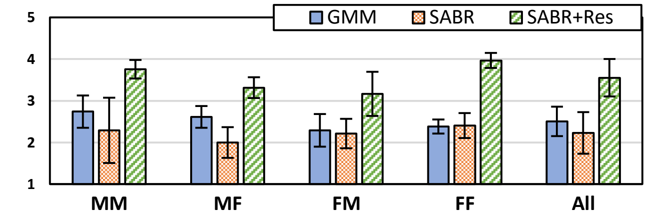

Figure 4: MOS comparison for GMM, SABR, and SABR+Res. Shown is the average performance of conversion for each possible conversion direction, and an aggregation over all directions.

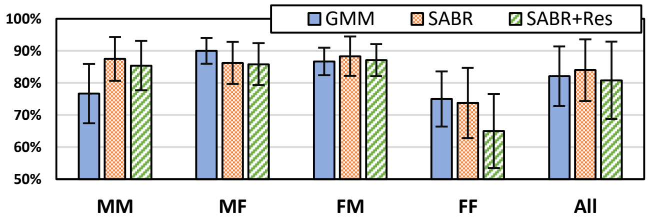

Figure 5: XAB identification rate for GMM, SABR, and SABR+Res VC methods. In all conditions except F-F, SABR+Res performed at least as well as the GMM condition. In the F-F case, the two speakers had a similar identity before conversion, making identifying them after conversion more difficult.

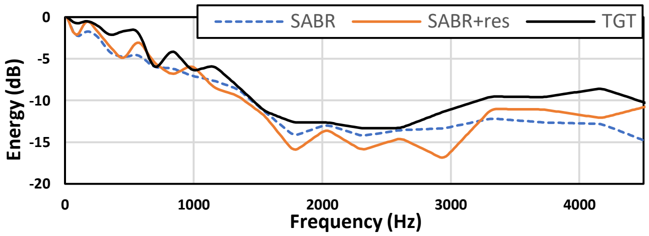

Figure 6: single frame comparison between SABR, SABR+Res, and target spectral envelopes. SABR+Res has substantially more detail in the spectrum as opposed to the original SABR spectrogram. The angular features in the latter half of the envelopes are from MFCC compression.

Currently, SABR anchors are built from phoneme centroids. As we reported previously [5, 20], SABR weights show less stability in turbulent and non-continuant segments. Thus, future developments will focus on modifying SABR to handle temporal anchors (e.g. using Tibshirani's Fused LASSO [21]). Additionally, increasing the number of anchors by taking into account allophones may improve the performance of SABR, even in extremely limited data conditions.

7. ACKNOWLEDGEMENTS

We would like to thank Profs. John Levis and Evgeny Chukarev-Hudialainen for their assistance. This work was partially supported by NSF Grants 1619212 and 1623750.

[1] Audio samples of SABR, SABR+Res, and GMM VC are available at http://people.tamu.edu/~cliberatore/samples/sabr.icassp2018.html

8. REFERENCES

[1] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, no. 4, pp. 1313-1323, 2012.

[2] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication,* vol. 51, no. 10, pp. 920-932, 2009.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing,* vol. 6, no. 2, pp. 131-142, 1998.

[4] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *Speech Synthesis Workshop*, 2013, pp. 201-206.

[5] C. Liberatore, S. Aryal, Z. Wang, S. Polsley, and R. Gutierrez-Osuna, "SABR: sparse, anchor-based representation of the speech signal," in *Interspeech*, 2015, pp. 608-612.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society,* vol. Series B, pp. 267-288, 1996.

[7] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004, pp. 223-224.

[8] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 8, pp. 2222-2235, 2007.

[9] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 3, pp. 556-566, 2013.

[10] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," in *ICASSP*, 2014, pp. 7894-7898: IEEE.

[11] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech and Language Processing,* vol. 22, no. 10, pp. 1506-1521, 2014.

[12] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics,* vol. 32, no. 2, pp. 407-499, 2004.

[13] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE transactions on audio, speech, and language processing,* vol. 20, no. 3, pp. 806-817, 2012.

[14] 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.

[15] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Transactions on Speech and Audio Processing,* vol. 13, no. 5, pp. 930-944, 2005.

[16] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, no. 5, pp. 922-931, 2010.

[17] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology,* vol. 27, no. 6, pp. 349-353, 2006.

[18] S. Buchholz and J. Latorre, "Crowdsourcing Preference Tests, and How to Detect Cheating," in *Interspeech*, 2011, pp. 3053-3056.

[19] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *ICASSP*, 2001, vol. 2, pp. 841-844: IEEE.

[20] C. Liberatore and R. Gutierrez-Osuna, "Generating Gestural Scores from Acoustics Through a Sparse Anchor-Based Representation of Speech," in *Interspeech*, 2016, pp. 1507-1511.

[21] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 67, no. 1, pp. 91-108, 2005.