EXEMPLAR SELECTION METHODS IN VOICE CONVERSION

Guanlong Zhao and Ricardo Gutierrez-Osuna[1]

Department of Computer Science and Engineering, Texas A&M University, USA
{gzhao, rgutier}@tamu.edu

ABSTRACT

Exemplar-based methods for voice conversion often use a large number of randomly-selected exemplars to ensure good coverage. As a result, the factorization step can be costly. This paper presents two algorithms that can be used to construct compact sets of exemplars. The first algorithm uses a forward selection procedure to build the exemplar set sequentially, selecting exemplar pairs that minimize the joint reconstruction error on source and target frames. The second algorithm uses a backward elimination procedure to remove exemplars that contribute the least to the factorization. We evaluate both selection strategies on voice conversion tasks using the ARCTIC corpus. Our results using objective measures and subjective listening tests show that both strategies can significantly reduce the size of the exemplar set (five-fold, in our experiments) while achieving the same performance on voice conversion.

Index Terms**—**voice conversion, sparse reconstruction, exemplar selection

1. INTRODUCTION

Voice conversion (VC) aims to convert utterances from a source speaker to sound as if a target speaker had produced them. VC has a number of real-world applications, from building personalized text-to-speech synthesizers [1] to improving speaker spoofing systems [2]. A number of VC techniques have been proposed over the years, statistical mappings and frequency warping being among the most common. A new VC framework based on exemplars and non-negative matrix factorization (NMF) has recently emerged [3-6]. In this "exemplar-based" framework, the speech signal is decomposed into a sparse, non-negative weight matrix and a set of preselected exemplars. These exemplars are generally acoustic frames from the source and target speaker's speech. VC is performed by applying NMF to the source utterance, and then combining the source weight matrix with the target exemplars; see Fig. 1.

Exemplar-based methods have several advantages over statistical methods, e.g., based on Gaussian mixture models (GMM). Takashima et al. [3] have shown that exemplar-based methods can produce more natural and higher quality speech than GMMs, while Wu et al. [5] have shown that they require smaller corpora than conventional methods [7]. However, exemplar-based methods select exemplars randomly [3-6], so the exemplar set tends to be large (from 1,000s to 10,000s) to ensure good coverage. As a result, the factorization step becomes time consuming; NMF is NP-hard, though polynomial-time approximations exist [8].

To address this issue, this paper describes two exemplar selection algorithms that can be used to build a compact exemplar set. The first algorithm operates in a forward fashion, incrementally adding new exemplars to the set in order to minimize the *joint* reconstruction error on source and target frames. The second algorithm operates in a backward fashion: starting from a large exemplar set, it sequentially removes exemplars that contribute the least to the factorization. Through a series of objective tests (Mel-cepstral distortion) and subjective listening tests, we show that the proposed algorithms can provide a more compact exemplar set without sacrificing VC performance.

Relation to prior work. Two prior studies have addressed the problem of building a compact exemplar set. Aihara et al. [9] optimize the exemplar set using an NMF variant that relaxes negativity constraints and a fast solver based on the alternating direction method of multipliers. In their study, an initial set of 5,000 exemplars was trimmed down to 1,000 exemplars without significant loss of quality. According to our results, further reductions in size (down to 200 exemplars) are also possible without introducing noticeable distortions. Liberatore et al. [10] use phoneme centroids as exemplars, referred to as "anchors" in their original paper, so their approach automatically leads to a compact exemplar set (i.e., the size of the phoneme set). In contrast, our approach does not require phonetic labels, and avoids averaging out spectral detail.

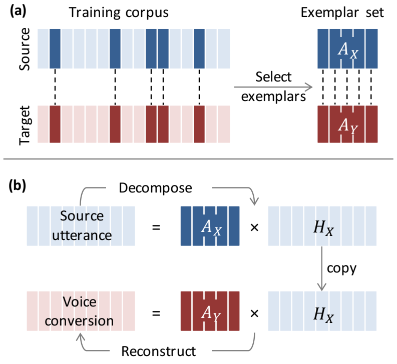

Fig. 1. Exemplar-based voice conversion: (a) exemplar selection during training; (b) voice conversion through NMF

[1] Work supported by NSF awards 1619212 and 1623750.

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

2. LITERATURE REVIEW

Much of the research on VC has focused on two approaches: statistical mappings and frequency warping. A conventional statistical approach is to model the joint distribution of source and target spectra with a GMM [1, 7], then map source spectra into target spectra using MMSE or ML criteria. Other statistical models, such as neural networks [11], partial least squares [12] and HMMs [13] have also been used with success. A second major approach consists of performing frequency warping to "align" the spectral power of the two speakers [14]. Frequency warping can generate speech of high acoustic quality, though the voice conversions tend to be less convincing. Hybrid approaches, which use GMMs to model the distribution of local frequency warps, have shown to achieve good acoustic quality and voice conversion results [15].

Sparse representations have recently garnered much attention as an alternative approach to VC. Takashima et al. [3] first used an exemplar-based sparse method to tackle the problem of VC in noisy environments by including noise exemplars in addition to source and target exemplars. Wu et al. [5] refined the sparse representation by jointly estimating the weight matrix using low- and high-resolution features to preserve temporal structure and spectral details. Aihara et al. [16] extended the exemplar-based method to perform many-to-many VC using dictionaries from a pool of speakers.

3. CONVENTIONAL NMF-BASED VC

Let $X \in R^{D \times T}$ denote a source utterance of length T, each frame represented by a D-dimensional vector of non-negative features (e.g., STFT spectra). Given an exemplar set $A_X \in R^{D \times N}$ containing N exemplars from the source speaker, NMF decomposes X as follows,

$$X \approx A_X H \tag{1}$$

where $H \in R^{N \times T}$ is a weight matrix constrained to be non-negative and sparse. H can be approximated by minimizing the objective function,

$$d_{KL}(X, A_X H) + \lambda \|H\|_1, \qquad subject\ to\ H \geq 0 \tag{2}$$

$d_{KL}(\cdot)$ is the KL-divergence, and λ is a sparsity parameter. To generate a target utterance $\hat{Y} \in R^{D \times T}$ with the same linguistic content as X, we replace A_X with an exemplar set $A_Y \in R^{D \times N}$ from the target speaker, and recombine with H:

$$\hat{Y} = A_Y H \tag{3}$$

As a result, $\hat{Y}$ will have the identity of the target speaker and the linguistic content of the source speaker.

3.1. Joint NMF for voice conversion

Using high-resolution features (e.g., STFT spectra) allows NMF to achieve high acoustic quality and naturalness. Further improvements can also be achieved if exemplars contain multiple (consecutive) frames to capture contextual information. Unfortunately, using multi-frame *and* high-resolution features makes NMF computationally expensive. To address this issue, Wu et al. [5] have recently proposed to combine single-frame high-res features with multi-frame low-res features. The approach consists of factorizing high-res (X^{HR}) and low-res (X^{LR}) features as:

$$X^{HR} \approx A_X^{HR} H\,; \quad X^{LR} \approx A_X^{LR} H \tag{4}$$

by minimizing the joint cost function:

$$\alpha d_{KL}(X^{HR}, A_X^{HR} H) + (1-\alpha) d_{KL}(X^{LR}, A_X^{LR} H) + \lambda \|H\|_1 \tag{5}$$

where α controls the contribution of low-res and high-res features. Once H is found (see [5] for details), the target utterance is then reconstructed using *only* high-res features:

$$\hat{Y}^{HR} = A_Y^{HR} H \tag{6}$$

In our study, we use the 513-dim STRAIGHT spectrogram [17] as high-res features, and 23 Mel-scale filter bank energies as low-res features.

4. EXEMPLAR SELECTION

Compiling an exemplar set starts by time-aligning source and target utterances in a training corpus. Once aligned, a subset of exemplar pairs is randomly selected to form matrices $\{(A_X^{HR}, A_X^{LR}), (A_Y^{HR}, A_Y^{LR})\}$. To ensure good coverage of the acoustic space, this subset contains a large number of exemplars (1,000s to 10,000s) [3-6]. However, a large exemplar subset makes computation of matrix H very expensive. In addition, random selection can inadvertently select pairs of exemplars that are acoustically mismatched due to alignment errors. To address these issues, we propose two algorithms that select a compact subset while avoiding misaligned pairs.

4.1. Forward selection

Our first algorithm uses a forward selection procedure to build the exemplar subset. 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; this ensures that misaligned frames are avoided. The algorithm is summarized in Table 1. Using NMF to perform the decomposition becomes prohibitive. For this reason, during forward selection we replace NMF with the pseudo-inverse solution:

$$H_X = \left(A_X'^{\,T} A_X'\right) A_X'^{\,T} X_D; \quad H_Y = \left(A_Y'^{\,T} A_Y'\right) A_Y'^{\,T} Y_D \tag{7}$$

which can be computed efficiently with orthogonal least squares (OLS) [18].

4.2. Backward elimination

Our second algorithm uses backward elimination to obtain a compact subset. It is based on the observation that, due to the NMF sparsity constraints, only a small portion of exemplars are active at any one time. The algorithm starts with a large subset of exemplars, and sequentially removes those exemplars with the lowest activation weights in NMF,

Table 1. Pseudocode for forward selection

```
**Inputs**: training set $(X_T, Y_T)$, development set $(X_D, Y_D)$, # exemplars $N$

$A_X = \{\emptyset\}, A_Y = \{\emptyset\}$ % initialize exemplar set
for i = 1:N % add a new exemplar to $(A_X, A_Y)$
  for j = 1:T % for each exemplar candidate in the training set $\{x_T^j, y_T^j\}$
      $A'_X = A_X \cup \{x_T^j\},\ A'_Y = A_Y \cup \{y_T^j\}$ % add exemplar to $(A_X, A_Y)$
      $X_D \approx A'_X H_X;\ \ Y_D \approx A'_Y H_Y$ % decompose development set
      $\epsilon^j = w(X_D - A'_X H_X) + (1-w)(Y_D - A'_Y H_Y)$ % compute error
  $k = \underset{j}{\operatorname{argmin}}\, \epsilon^j$ % select exemplar pair with lowest error
  $A_X = A_X \cup \{x_T^k\}, A_Y = A_Y \cup \{y_T^k\}$ % add exemplar $k$ to $(A_X, A_Y)$
return $A_X, A_Y$
```

Table 2. Pseudocode for backward elimination

```
**Inputs**: initial exemplar set $(A_X, A_Y)$, dev. set $(X_D, Y_D)$, # exemplars $N'$

while ($|A_X| > N'$) % $(A_X, A_Y)$ have more than $N'$ exemplars
    $X_D \approx A_X H_X, Y_D \approx A_Y H_Y$ % decompose development set
    % compute avg. activation of each exemplar (i.e., rows in $H_X, H_Y$):
    $\overline{H}_X = \text{mean}(H_X), \overline{H}_Y = \text{mean}(H_Y)$
    % sort exemplars based on their source and target activation:
    $s = sort\ (\overline{H}_X + \overline{H}_Y)$
    % remove the lowest $\eta$% exemplars (denoted by $k_{LOW}$)
    $A_X = A_X - A_X[k_{LOW}], A_Y = A_Y - A_Y[k_{LOW}]$
return $A_X, A_Y$
```

computed on a development set. The algorithm is summarized in Table 2. It requires selecting a parameter η that determines how many (as a percentage) of the least active exemplars are removed at each step. Large η speed up the process, at the expense of eliminating exemplars that may become critical at a later time. Compared to the forward selection algorithm, backward elimination only requires a small number of decompositions since the size of the exemplar set decays geometrically. As an example, using $\eta = 20\%$ an initial set of 10,000 exemplars can be reduced to 44 exemplars (i.e., the number of phonemes in English) in 25 steps. For this reason, the backward elimination procedure can afford to use NMF during the decomposition step.

5. EXPERIMENTAL SETUP

We evaluated the two algorithms on four speakers from ARCTIC [19]: two males (BDL, RMS) and two females (SLT and CLB). For each speaker, we selected three sets of utterances, 10 for training, 10 for development, and 50 for testing. Our choice for such small training and development sets was motivated by applications where collecting a large corpus is impractical (e.g., pronunciation training [20, 21]). To ensure good phonetic balance, training and development utterances were selected using a maximum entropy criterion.

We extracted high-res spectra with STRAIGHT [17] (25ms window, 5ms shift). To generate low-res features, we passed STRAIGHT spectra through a 23-dim Mel-scale filter bank. We also used STRAIGHT spectra to compute 25 MFCCs as the representation for time-alignment, OLS, and initialization of the backward elimination procedure. Based on preliminary experiments, parameters α (balance of low-res and high-res features) and λ (sparsity) were set to 0.1 and 0.7, respectively, and the maximum number of NMF iterations was set to 300. Our preliminary experiments revealed no major differences in acoustic quality when using multiple frames (see section 3.1); for this reason, we used a single HR and a single LR frame (i.e., no context).

To perform VC, we transformed source into target spectra as described in section 3, and normalized the source F_0 contour to match target speaker's range using the log-scale mean and variance normalization method in [7]. Then, we used STRAIGHT to synthesize speech from the transformed spectra, normalized F_0 contour, and the source aperiodicity signal. We tested three VC implementations[2]:

- **RAND (baseline)**: Time-aligned training and dev sets using DTW, selected exemplar pairs randomly, and used joint NMF to obtain the weight matrix.
- **FWD**: Used forward selection (Table 1) to find a compact exemplar set; all other settings as in RAND
- **BKW**: Used backward elimination (Table 2) to find a compact exemplar set; other settings as in RAND.

We evaluated each system on four VC tasks: BDL to RMS (m-m), CLB to SLT (f-f), BDL to SLT (m-f), and CLB to RMS (f-m). For RAND, we ran each VC pair 8 times and report the average result. For each task, we used up to 1,000 exemplars; preliminary experiments showed only marginal reductions in MCD beyond that number.

6. RESULTS

6.1. Objective evaluation

We evaluated forward selection as a function of the number of exemplars selected. For each VC task, we used the optimal w value and then computed MCDs. Fig. 2a summarizes results, averaged over the 4 tasks. We observe that FWD consistently has a lower MCD than RAND. More interestingly, FWD can achieve a similar MCD as RAND using a fraction of the exemplars. For example, RAND performs best when it has 1,000 exemplars (MCD: 2.29), whereas FWD only needs 300 exemplars to achieve a similar result (MCD: 2.23).

Next, we evaluated backward elimination as a function of the cutoff threshold η. To obtain the initial exemplar set (size=1,000), we applied hierarchical clustering to time-aligned source and target training frames, trimmed the dendrogram when it reached 1,000 clusters, and used the paired centroid frames of each cluster as the initial set. Average results over the four VC tasks are shown Fig. 3. The MCD follows a similar trend regardless of the cutoff threshold, and remains nearly unchanged as the number of exemplars is reduced from 1000 to 250. The MCD reaches a minimum at 100 exemplars and then increases for smaller

[2] We also tried k-means centroids of training data as exemplars, but do not report the results since they were uninteresting.

exemplar subsets – notice that the difference between the highest and lowest MCD is rather small (0.03). This result shows that using η =50% we can reduce the exemplar set four-fold (from 1000 to 250) in only 2 iterations with no loss in acoustic quality. A possible explanation is that only a small portion of exemplars contribute to the NMF activation weights, which allows us to use an aggressive cutoff threshold without removing critical exemplars.

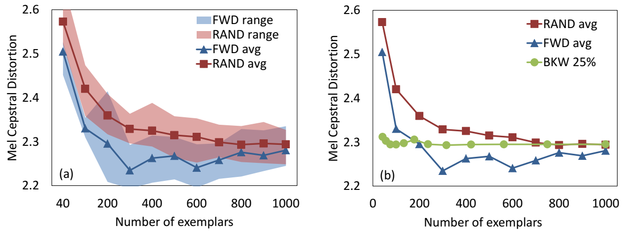

Fig. 2. (a) MCD for forward and random selection; shaded region indicates min-max range. (b) Average MCD for the three methods

Lastly, we evaluated the two algorithms against random selection. For BKW, since all cutoff thresholds perform similarly, we use $\eta = 25\%$ because it gives the closest number of data points (i.e. 12) as the other two methods (i.e. 11). The comparison is shown in Fig. 2b. FWD achieves the lowest MCD among the three methods, whereas BKW performs best with limited exemplar set size.

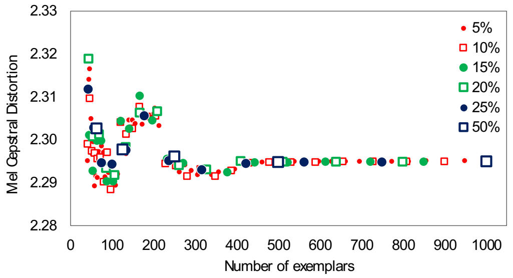

Fig. 3. MCD during backward elimination for different cutoff thresholds ($\eta = 5, 10, 15, 20, 25, 50\%$)*, averaged over 4 pairs*

6.2. Subjective evaluation

In a final experiment, we evaluated the three methods through a set of subjective listening tests. Following prior studies [22], we used a 5-point mean opinion score (MOS) to rate acoustic quality, and a voice similarity score (VSS) ranging from -7 (definitely different speakers) to +7 (definitely same speaker) to rate speaker identity. For the audio quality test, we used 1000, 200, and 204 exemplars for RAND, FWD, and BKW, respectively, since they had similar MCDs. Sixteen participants rated 120 VC utterances: 40 utts per algorithm, 10 utts per speaker pair; all utts were randomly ordered. We found no significant differences between the three methods –see Fig. 4a; thus, the two proposed algorithms can achieve the same acoustic quality as random selection with 20% of the exemplars.

For the speaker identity test, 14 participants rated 120 utterance pairs: 40 pairs (20 VC-source and 20 VC-target pairs) per system, 10 pairs per speaker pair. For each utt pair, participants were asked to first decide whether or not the two utts were from the same speaker, and then rate their confidence level on a 7-point scale. Utt pairs were presented in a random order. Source and target utts were resynthesized using NMF to avoid bias on speech quality. Following [22], responses were compiled into a VSS score. As shown in Fig. 4b, the three methods have similar VSS. Participants were "*quite a bit*" confident that VC and source utts were from different speakers (VSS≈4) but not sure if VC and target utts were from the same speaker (VSS ≈ 0.5). The low similarity between VC and target is caused by cross-gender conversions, which are rated as VSS=−4.5 ("different speakers"), whereas same-gender conversions are rated as VSS=5.3 ("same speaker"). One possible explanation is that prosody carries speaker identity information, and prosody differences are more distinguishable across genders. As a result, cross-gender VC utts have similar prosody as the source speaker, which participants use –in addition to spectral cues—to make their decisions. Nevertheless, the three systems perform similarly, with the two proposed methods using 5 times fewer exemplars than the baseline.

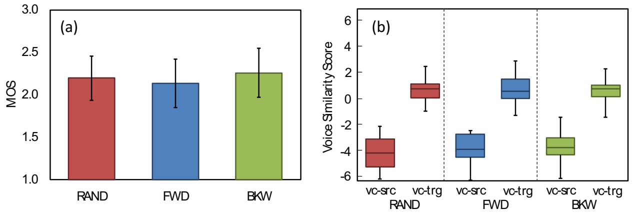

Fig. 4. (a) Speech quality results with 95% confidence interval. (b) Speaker identity results (vc-source: VSS between voice conversion and source speaker)

7. CONCLUSION AND FUTURE WORK

Exemplar-based VC methods often require a large set of exemplars to achieve good performance. In this paper, we have proposed two complementary strategies that can be used to construct compact exemplar sets. The first strategy (forward selection) builds a compact set by selecting exemplar pairs that reduce the joint reconstruction error of source and target utterances. The second strategy (backward elimination) excludes exemplars that contribute the least to the factorization. Objective and subjective measures show that both strategies can significantly reduce the number of exemplars (from 1000 to 200, in our experiments), without sacrificing VC performance. Using a contemporary PC, both algorithms take ~10-15 min to identify 200 exemplars.

For computational reasons, the forward procedure uses least-squares as a proxy of NMF, yet achieves lower NMF reconstruction error than the backward procedure, which does use NMF to exclude frames. Future work will examine ways in which both procedures can be combined, e.g. using exemplars generated by forward selection as the initial set for backward elimination. An additional direction of future work is to improve the computational efficiency of the forward selection algorithm, e.g., using heuristic rules to reduce the search space or the alternating direction method of multipliers [9].

REFERENCES

[1] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *ICASSP*, 1998, pp. 285-288.

[2] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *ICASSP*, 2012, pp. 4401-4404.

[3] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *IEEE Spoken Language Technology Workshop*, 2012, pp. 313-317.

[4] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Noise-robust voice conversion based on sparse spectral mapping using non-negative matrix factorization," *IEICE Transactions on Information and Systems,* vol. 97, pp. 1411-1418, 2014.

[5] Z. Wu, E. S. Chng, and H. Li, "Exemplar-based voice conversion using joint nonnegative matrix factorization," *Multimedia Tools and Applications,* vol. 74, pp. 9943-9958, 2015.

[6] H. Ming, D. Huang, L. Xie, S. Zhang, M. Dong, and H. Li, "Exemplar-based sparse representation of timbre and prosody for voice conversion," in *ICASSP*, 2016, pp. 5175-5179.

[7] 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, pp. 2222-2235, 2007.

[8] S. A. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM Journal on Optimization,* vol. 20, pp. 1364-1377, 2009.

[9] R. Aihara, T. Takiguchi, and Y. Ariki, "Semi-non-negative matrix factorization using alternating direction method of multipliers for voice conversion," in *ICASSP*, 2016, pp. 5170-5174.

[10] 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.

[11] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral Mapping Using Artificial Neural Networks for Voice Conversion," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, pp. 954-964, 2010.

[12] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, pp. 912-921, 2010.

[13] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *ICASSP*, 1996, pp. 389-392.

[14] D. Sundermann and H. Ney, "VTLN-based voice conversion," in *The Third IEEE International Symposium on Signal Processing and Information Technology*, 2003, pp. 556-559.

[15] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, pp. 922-931, 2010.

[16] R. Aihara, T. Takiguchi, and Y. Ariki, "Many-to-many Voice Conversion Based on Multiple Non-negative Matrix Factorization," in *Interspeech*, 2015, pp. 2749-2753.

[17] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication,* vol. 27, pp. 187-207, 1999.

[18] S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal Least-Squares Learning Algorithm for Radial Basis Function Networks," *IEEE Transactions on Neural Networks,* vol. 2, pp. 302-309, 1991.

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

[20] S. Aryal, D. Felps, and R. Gutierrez-Osuna, "Foreign accent conversion through voice morphing," in *Interspeech*, 2013, pp. 3077-3081.

[21] S. Aryal and R. Gutierrez-Osuna, "Can Voice Conversion Be Used to Reduce Non-Native Accents?," in *ICASSP*, 2014, pp. 7879-7883.

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