Application of Voice Conversion for Cross-Language Rap Singing Transformation

Oytun Türk[1], Osman Büyük[2,3], Ali Haznedaroglu[2,3], Levent M. Arslan[2,3]

Speech Group, DFKI GmbH Language Technology Lab, Berlin, Germany
Sestek Inc., ITU Ayazaga Kampusu, ARI-1 Teknopark Binasi, Maslak, Istanbul, Turkey
Electrical and Electronics Eng. Dept., Bogazici University, Istanbul, Turkey
<oytun.turk>@dfki.de, <osman.buyuk, ali.haznedaroglu, levent.arslan>@sestek.com.tr

Abstract

Voice conversion enables generation of a desired speaker's voice from audio recordings of another speaker. In this paper, we focus on a music application and describe the first steps towards generating voices of music celebrities using conventional voice conversion techniques. Specifically, rap singing transformations from English to Spanish are performed using parallel training material in English. Weighted codebook mapping based voice conversion with two different alignment methods and temporal smoothing of the transformation filter are employed. The first aligner uses a HMM trained for each source recording to force-align the corresponding target recording. The second aligner employs speaker-independent HMMs trained from a large number of speakers. Additionally, a smoothing step is devised to reduce discontinuities and to improve performance. The results of subjective evaluations indicate that both aligners perform equivalently well. The proposed smoothing technique improves both similarity to target singer and quality significantly regardless of the alignment method.

Index Terms: voice conversion, singing voice transformation, weighted codebook mapping

(*) Online demo available at http://www.voxonic.com

1. Introduction

Voice conversion provides a useful framework for generating a desired speaker's voice from audio recordings of another speaker. In cross-language domain, it can be used for producing the voice of a speaker in a language that s/he can not speak. In this study, we focus on a music application, an area which has not yet been commonly exploited by voice conversion researchers. We describe the first steps towards generating voices of music celebrities using conventional voice conversion techniques. Specifically, rap singing transformations from English to Spanish are performed using parallel training material in English as shown in Figure 1. We evaluate the performance of a conventional voice conversion method for the given task.

Voice conversion algorithms have been commonly used in speech synthesis domain to create new voices from existing ones [1, 2], to generate multi-lingual voices [3] or a specific expressive style [4]. In [5], a voice conversion algorithm is proposed to recover voices of patients with laryngectomy (larynx-removal). Further application areas include second language learning for providing feedback to the learner, video gaming for developing adaptive voices, international movie dubbing, Karaoke singing (for example as in [6]), and speech and speaker recognition.

Training and transformation constitute the two major steps of voice conversion. In training, source and target speaker acoustic characteristics are mapped using feature extraction, alignment, and machine learning stages. Spectral characteristics are commonly modeled using linear prediction coefficients (LPCs), line spectral frequencies (LSFs), mel frequency cepstral coefficients (MFCCs), or sinusoidal model parameters. Duration, energy, and f0 provide the additional information required for prosody mapping and transformation. Codebook mapping [7], weighted codebook mapping [8], Gaussian Mixture Models (GMMs) [1, 9], and Hidden Markov Models (HMMs) [10] have been popular approaches to learn the acoustic mapping between the source and the target speaker characteristics. In the transformation stage, acoustic features of source speaker are converted to match target characteristics using signal processing techniques. Time varying filtering and prosody modification using overlap-add methods, sinusoidal model based modifications, and homomorphic processing are commonly employed.

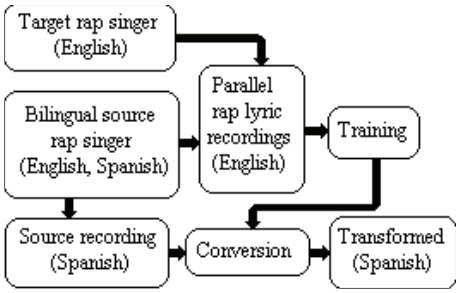

Figure 1. Flowchart of the cross-language rap singing transformation system.

Since alignment is an important step that affects voice conversion performance, we focus on automatic alignment methods based on HMMs in the first part of this work. Ideally, it would be required to collect rap recordings to train rap singing specific HMMs. To avoid the elaborate amount of work, we have chosen two alternative approaches that can provide automatic alignment without the requirement of text transcriptions: Sentence-HMM (SHMM) and Phonetic-HMM (PHMM). In SHMM, a HMM is trained for each source utterance and the corresponding target utterance is force-aligned with the source HMM. In PHMM, context independent HMMs trained from a large amount of multi-speaker recordings from Turkish are used for alignment. Alignment with context independent monophone models can be achieved with: (i) an open-loop phoneme recognizer, (ii) force alignment to previous recognition result, or (iii) known text depending on application. In our algorithm, an open-loop phoneme recognizer is used to find the best matching phoneme sequence to the source singer's training recording. Then, target singer's parallel recording is force aligned to the recognized phoneme string. We have also tested PHMMs trained over a smaller corpus in American English with degraded performance. Therefore, Turkish HMMs are used in this study although training HMMs from an equivalent size corpus in English is likely to improve results.

In addition to comparing two alignment methods in a rap singing voice conversion task, a temporal smoothing procedure is devised to reduce discontinuities in conversion output especially observed at phoneme boundaries. Conversion is applied in two passes to smooth the vocal tract transformation filters. In the first pass, the DFT of the impulse response of the vocal tract transformation filter is estimated for each speech frame. Then, each frequency bin is smoothed by weighting with bins from neighboring frames using a Hanning window.

Section 2 describes the cross-language rap singing transformation application and the available database in detail. We review weighted codebook mapping highlighting the improvements proposed in this paper. The details of the two aligners and temporal smoothing of the transformation filter are described next. In Section 3, a subjective listening test is presented for evaluating similarity to target singer and quality. Finally, results and steps towards future research are discussed in Section 4.

2. Method

2.1. Target Application and Database

The main purpose of the rap singing voice conversion task is to transform the voice of an amateur rap singer (source) into that of a celebrity singer (target). The target singer is a native American English speaker. The source singer is a bilingual speaker of American English and Spanish. The training database used for mapping the source and target acoustic spaces is recorded in English language whereas the transformation database is in Spanish. Therefore, the aim is to have the target singer sing rap in a language which he can not speak. The transformation outputs are mixed with music to get the final form of the rap song in Spanish. The transformation is formulated by a training phase in which the acoustic characteristics of the source and target singers are modeled and mapped using a weighted codebook mapping approach [8].

A total of 136 short English rap song segments sampled at 44.1 kHz available from the target rap singer are used as the training material. Total duration of this training database is approximately 4.5 minutes. The duration of the segments vary from 1 to 4 seconds. Since we only had access to pre-recorded target training material, any optimizations to be realized during database collection needed to be carried out on source speaker's side.

Source singer recordings are collected in a professional studio with high quality equipment. The source singer is directed to mimic the singing style of the target celebrity singer in order to reduce the amount of prosody modifications required and to enhance automatic alignment accuracy. Each target segment is repeatedly recorded and best stylistic match is selected by a sound engineer during database collection.

2.2. Weighted Codebook Mapping

Spectral envelope is transformed using the weighted codebook mapping algorithm STASC [8] with the improvements proposed in [11] using a parallel training database. The training algorithm extracts average source and target LSF vectors after an LP analysis of order 46 for 44.1 KHz recordings using a skip rate of 10 msec. and a window size of 25 msec. Pre-emphasis with a digital filter of the form $1\text{-}0.97z^{-1}$ was applied to enhance LP modeling at higher frequencies. Duration, average energy, and average f0 of voiced frames for each source and target phoneme pair are also extracted. In order to eliminate source-target pairs that originate from misaligned labels, we use an automatic outlier detection method [11]. In this method, confidence measures based on the distributions of LSF, f0, RMS energy, and duration differences between source and target codebook entries are computed. A single Gaussian is fitted to each distribution and the entries that have larger difference from the Gaussian mean by some factor of the standard deviation (typically 1.5 to 2.5) are eliminated. This helps to exclude relatively different source and target pairs from the codebooks, resulting in a more stable and less discontinuous transformation function.

In the transformation stage, source LSF vectors for each input speech frame are matched with source codebook entries using an inverse harmonic LSF weighting based distance measure. The estimated target LSF vector is computed as a weighted average of the best target codebook matches (typically 3 to 10) using the inverse of the source distance values as weighting factors. Additionally, input spectral envelope is obtained as a weighted average of the source codebook entries instead of direct LP analysis to provide an additional smoothing step in the transformation function. Finally, FD-PSOLA is applied to realize prosody modifications to match average target pitch and speaking rate in parallel with vocal tract transformation with the estimated transformation filter. Details of the overall procedure can be found in [8, 11].

2.3. Alignment

2.3.1. Sentence-HMM (SHMM)

SHMM aligner aims to fit an acoustic model for each source singer recording and then to perform alignment between each source-target singer recording pair using the model as the reference. This method does not require the phonetic translation of the orthographic transcription for the training utterances, however it assumes that both source and target singers are singing the same lyrics during the training session. The method is implemented as follows. Silence regions at the beginning and end of each recording are removed using an energy based end-point detection algorithm. RMS energy normalization is applied to each utterance in order to normalize recording gain level differences. Next, cepstrum coefficients are extracted along with log-energy and probability of voicing parameters on a frame-by-frame basis for each recording. Cepstral mean normalization is applied in order to account for possible microphone differences during source and target singer recordings. SHMMs are trained based on the parameter vector sequences for each recording of the source singer. The number of states for each recording is set proportional to its duration. Segmental K-Means algorithm is applied first during training in order to initialize the state means. This is followed by several iterations of the Baum-Welch algorithm until convergence is achieved. The covariance matrix is estimated over the whole utterance parameter vector once, and is not updated during training due to limited amount of data. Finally, best state sequence is calculated using Viterbi algorithm for source-target recording pairs.

2.3.2. Phonetic-HMM (PHMM)

PHMM alignment is done using phone-based HMMs trained from 53 hours of speech database in Turkish. The database was collected at 16kHz, 16 bits at different Turkish universities [12, 13, 14] and at Sestek. Context-independent HMMs for 29 phones in Turkish language plus two silences and one short pause models are trained using recordings of

various speakers. Speakers in database constituted nearly equal gender distributions and read phonetically balanced sentences. 5 states and 10 Gaussian mixtures are used for each speaker independent monophone models. HTK 3.2 is used for HMM training [15].

In our voice conversion experiments, training recordings of source singer in English are recognized with an open-loop phoneme recognition network to find the best matching phoneme sequence to the recording. Then, the parallel recording of the target singer is force aligned to the recognized phoneme string. Transformation utterances of source singer in Spanish are also recognized using the same open-loop phoneme recognizer. In this alignment procedure, recognition errors during phoneme recognition network may cause alignment inaccuracies. However, if the texts of recordings are also available, these can be used to get rid of recognition errors and the corresponding alignment problems. In the given task, we did not have access to the lyrics of the rap recordings. Therefore, we have not used "alignment-to-text" option at all.

2.4. Temporal Smoothing

Temporal smoothing of the vocal tract transformation filter is applied in two passes. In the first pass, the DFT of the impulse response of the vocal tract transformation filter is calculated using LP analysis and target LSF estimation as described in Section 2.2. To reduce discontinuities due to abrupt changes or discontinuities across frequency bins in neighboring speech frames, each frequency bin is temporally smoothed by using a Hanning window and corresponding frequency bins from neighboring speech frames in a second pass. Choice of the number of neighboring frames plays an important role in shaping the trade-off between voice conversion quality and similarity. Although increasing number of neighboring frames can increase output quality, it may have an adverse effect on similarity to target. After informal evaluations, the best choice for number of neighboring frames is found to be 4 in voice conversion tests reported in this paper. Therefore, the DFTs of vocal tract transformation filter impulse responses for 4 previous, 4 next, and current speech frame are used in the smoothing procedure.

3. Evaluations and Results

3.1. Subjective Listening Test

In order to compare the performances of SHMM and PHMM aligners on rap singing voice conversion, 6 Spanish rap song segments are transformed using the resulting codebooks of the two alignment methods. The transformations are performed twice, once without temporal smoothing and once with smoothing. Therefore, our test consists of 4 sets of transformations:

- SHMM0: Sentence-HMM aligner, no smoothing
- SHMM1: Sentence-HMM aligner, with smoothing
- PHMM0: Phonetic-HMM aligner, no smoothing
- PHMM1: Phonetic-HMM aligner, with smoothing

Transformations are evaluated by a subjective listening test conducted on 16 subjects who were researchers in speech technology. Two scoring criteria are used in the test:

- Similarity to target on an increasing similarity scale from 1 to 10
- Quality using the standard 5-point MOS scale

First, each transformation set is paired with a target voice recording, resulting in 24 target-transform pairs. Then, 3 target-target and 3 source-target pairs are included in the test as well to serve as reference pairs, resulting in a total of 30 test pairs. Each of these pairs is evaluated by each subject in random order in terms of similarity to target and quality of the transformation output. The listening test was carried out in silent office environment with high quality headphones.

3.2. Results

Tables 1 and 2 show the average similarity and quality scores respectively for different cases. In Table 1, we observe that the performances of SHMM and PHMM aligners are close in terms of similarity to target, with PHMM performing slightly better. The source singer was also rated as slightly similar to target singer with a score of 3.52 out of 10 as he tried to mimic the target singer's style during training database collection. Tables 1 and 2 show that smoothing improves both similarity and quality for both SHMM and PHMM as verified by statistical tests. We think that the improvement in similarity to target with smoothing is due to the significant improvement in quality shown in Table 2. Smoothing also helps to eliminate abrupt changes in conversion filters around phoneme boundaries resulting in outputs free of artifacts.

Figure 2 shows an example without and with smoothing. From the spectrograms, we observe that the transition between phonemes /j/ and /o/ are significantly distorted without smoothing. Through informal listening, we have confirmed that the similarity to target is not much affected in steady portions of the phonemes while artifacts such as the one shown in Figure 2 are mostly eliminated using a 9-point Hanning window to smooth out frequency bins of the transformation filter.

	Source–Target	Target-Target	SHMM-Target	PHMM-Target
Similarity (out of 10)	3.52	8.85	5.69 (SHMM0=5.00, SHMM1=6.39)	5.74 (PHMM0=5.14, PHMM1=6.35)

Table 1. Average similarity to target.

	Target	No smoothing	With smoothing
Quality (out of 5)	4.66	2.25 (SHMM0=2.40, PHMM0=2.10)	4.13 (SHMM1=4.19, PHMM1=4.07)

Table 2. Average quality.

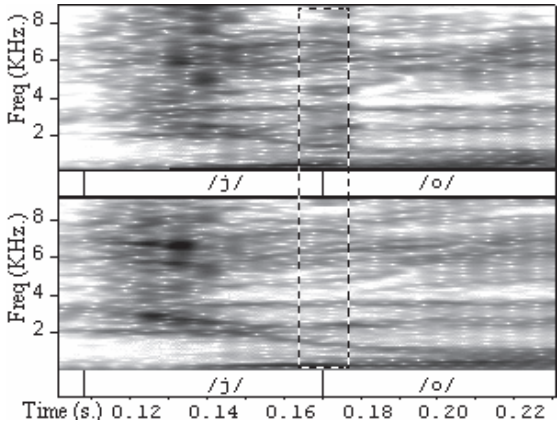

Figure 2. Transformation without (top) and with (bottom) smoothing.

Two-way analysis of variance (ANOVA) was performed to test how the two alignment methods and smoothing affect similarity to target and quality. It turned out that smoothing improves both similarity (p=0, F-ratio=45.82) and quality (p=0, F-ratio=568.04). The difference of similarity scores between SHMM and PHMM was not significant (p=0.7867, F-ratio=0.07). Quality scores for SHMM alignment were slightly better as compared to PHMM, the result being statistically significant (p=0.0104, F-ratio=6.63). We have not observed a combined effect of alignment method and smoothing on similarity (p=0.6651, F-ratio=0.19) and on quality (p=0.2624, F-ratio=1.26).

Figure 3 shows the median and the 25th and 75th percentiles of similarity and quality scores obtained for source-target, target-target-target, SHMM0, SHMM1, PHMM0, and PHMM1. The figures are sorted in increasing median value from left to right. Further investigations using pair-wise t-tests indicated significant differences for means of all pairs except "SHMM0 vs PHMM0" and "PHMM1 vs SHMM1" for a confidence level of 99%. Therefore, two alignment methods did not differ significantly in terms of similarity and quality scores.

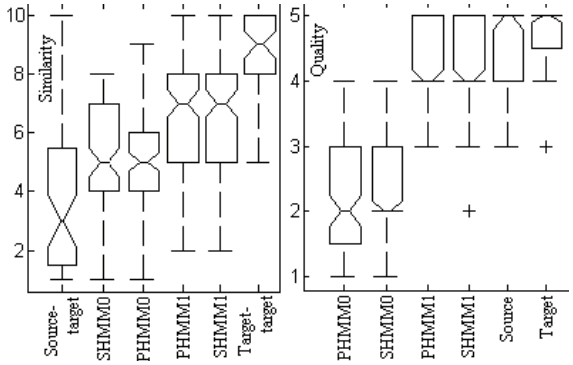

Figure 3. Left: Median, 25th, and 75th percentiles of similarity scores between source and target, target and target, and transformed and target using (SHMM0, SHMM1, PHMM0, PHMM1) methods. Right: Median, 25th, and 75th percentiles of quality scores for source, target, and transformed recordings.

4. Conclusions

In this study, we reported on the use of voice conversion for a cross-language rap singing transformation application. Subjective evaluations indicated that Phonetic-HMM based alignment resulted in slightly better similarity to target as compared to Sentence-HMM. Smoothing improved both similarity and quality scores significantly independent of the alignment method employed. Considering the automatic alignment framework employed to train phonetic HMMs, a relevant improvement is likely to be achieved if monophone HMMs trained using a large, phonetically-balanced, multi-speaker database in English had been employed for aligning the training recordings. Incorporating text information by using forced-alignment to text transcription can lead to additional improvements. In the ideal case, the phonetic HMMs should be trained using rap singing recordings with appropriate transcriptions. Since it could be hard to obtain sufficient amount of training data in rap singing domain, an alternative approach might be to adapt HMMs trained from large speech material to rap vocal recordings.

5. Acknowledgements

Oytun Türk's research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 211486 (SEMAINE). The remaining authors´ work has been supported by Devlet Planlama Teskilati DPT-TAM 2007K120610 and Tübitak TEYDEB 3070164 projects.

6. References

[1] Kain, A. and Macon, M., "Personalizing a speech synthesizer by voice adaptation", in Proc. of the Third ESCA/COCOSDA International Speech Synthesis Workshop 1998, pp. 225-230.

[2] Zhang, W., Shen, L. Q., and Tang, D., "Voice conversion based on acoustic feature transformation", in Proc. of the 6th National Conference on Man-Machine Speech Communications 2001.

[3] Latorre, J., Iwano, K., and Furui, S. "Polyglot synthesis using a mixture of monolingual corpora", in Proc. of the IEEE ICASSP 2005, vol.1, pp. 1-4.

[4] Türk, O. and Schroeder, M., "A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis", in Proc. of Interspeech 2008, pp. 2282-2285.

[5] Nakamura, K., Toda, T., Saruwatari, H., and Shikano, K., "Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech", in Proc. of the Interspeech 2006, Pittsburgh, Pennsylvania, pp. 1395-1398.

[6] Turajlic, E., Rentzos, D., Vaseghi, S., and Ching-Hsiang, H., "Evaluation of methods for parametric formant transformation in voice conversion", in Proc.of the IEEE ICASSP 2003, vol. 1, pp. 724-727.

[7] Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H., "Voice conversion through vector quantization", in Proc. of the IEEE ICASSP 1988, pp. 565-568.

[8] Arslan, L. M., "Speaker Transformation Algorithm using Segmental Codebooks", in Speech Communication, vol. 28 (1999), pp. 211-226.

[9] Stylianou, Y., Cappe, O., and Moulines, E., "Continuous probabilistic transform for voice conversion", in IEEE Trans. on Speech and Audio Proc., vol. 6 (1998), no. 2, pp. 131-142.

[10] Kim, E., Lee, S., and Oh, Y., "Hidden markov model based voice conversion using dynamic characteristics of speaker," in European Conference On Speech Communication And Technology 1997, pp. 1311–1314.

[11] Türk, O. and Arslan, L. M., "Robust processing techniques for voice conversion", in Computer Speech and Language 20 (2006), pp. 441-467.

[12] Erdogan, H., Buyuk, O., and Oflazer, K., "Incorporating language constraints in sub-word based speech recognition", in Proc. of the ASRU 2005, Cancun, Mexico.

[13] Salor, O., Pellom, B., Ciloglu, T., Hacioglu, K., and Demirekler, M., "On developing new text and audio corpora and speech recognition tools for the Turkish language", in Proc. of the ICSLP 2002.

[14] Arisoy, E., "Turkish dictation system for radiology and broadcast news applications", M.S. Thesis, Bogazici University, 2004.

[15] Young, S., Ollason, D., Valtchev, V., and Woodland, P., The HTK Book (for HTK Version 3.2), Entropic Cambridge Research Laboratory, 2002.