CROSS-GENDER VOICE CONVERSION WITH CONSTANT F0-RATIO AND AVERAGE BACKGROUND CONVERSION MODEL

Zbigniew Łatka[1], *Jakub Gałka*[1,2], *Bartosz Ziółko*[1,2]

[1]Techmo, Torfowa 1/5, 30-384 Kraków, Poland, techmo.pl
[2]AGH University of Science and Technology, Department of Electronics,
Al. Mickiewicza 30, 30-059 Kraków, Poland

ABSTRACT

This paper presents the method for spectral voice conversion using parallel training data. The proposed solution was submitted to the 2018 Voice Conversion Challenge. The method focuses on the preparation of the generative model for cross-gender voice conversion in differential-filtering framework. To improve the quality of the Gaussian mixture conversion model we introduced the usage of the averaged speaker background model pre-training step. Constant F_0 ratio transformation of source speech using WORLD vocoder was also proposed to improve cross-gender conversion quality. The evaluation results show that the proposed solution outperforms most of the concurrent systems submitted to the 2018 Voice Conversion Challenge, both in terms of speech quality and similarity. The system achieved 76% similarity score and 3.22 mean opinion score in cross-gender conversion task.

Index Terms— voice conversion, F_0 transformation, GMM, differential filtering.

1. INTRODUCTION

A term voice conversion (VC) describes the family of techniques dedicated for automatic speech modification, including conversion of nonlinguistic information such as voice timbre, prosody or pitch, while keeping the linguistic information unchanged. VC makes it potentially possible to synthesize various types of speech and speech sounds, even beyond human physical constraints [1]. A typical application of VC is speaker conversion [2]. The goal of this operation is to change the perceived identity of a speaker. To stimulate the development of voice conversion methods a 2016 [3] and 2018 Voice Conversion Challenges[1] (VCC) were organized by the international speech processing community. Both contests stimulated the interest in voice conversion research by providing the unified evaluation platform and required speech data. This paper presents our contribution to the 2018 VC Challenge.

A typical VC technique relies on statistical approach in which conversion function is trained using parallel speech data. Usually, the conversion process involves the alteration of spectral characteristics of speech signal as well as adjusting the excitation parameters. One of the most popular statistical VC methods is based on the maximum likelihood estimation of spectral parameter trajectory [4]. In this method, a Gaussian Mixture Model (GMM) of the probability of joint source-target features is employed for performing spectral conversion between the speakers. Similar to the statistical parametric speech synthesis frameworks [5], dynamic features statistics are used for computation of the converted spectrum sequence. To address the oversmoothing problem, additional features modeling, such as the global variance (GV), is incorporated. A typical VC system uses vocoding techniques to generate speech signal from the converted parameters sequences. This approach, however, causes significant audio quality degradation, even when sophisticated (eg. STRAIGHT [6]) vocoders are used.

To avoid this issue, a direct waveform modification technique, based on spectral differential filtering (DIFFVC), has been proposed [7]. Direct waveform filtering helps to improve the VC in terms of speech quality by removing the vocoder from the processing pipeline. Because spectral filtering affects only the timbre of voice, F_0 transformation has to be applied as part of audio preprocessing as well. Multifariousness of voice fundamental frequencies, produced by individual speakers, makes the choice of the F_0 transformation algorithm crucial for obtaining a high quality VC in different scenarios. Three main techniques were proposed earlier [7]:

- F_0 transformation using waveform modification,
- F_0 transformation using residual signal modification,
- F_0 transformation using STRAIGHT vocoder.

The first method uses waveform similarity-based overlap-add (WSOLA) [8] time-scaling algorithm followed by resampling to stretch or shrink the audio signal. While no signal parametrization is required in this method, it allows for achieving a high quality converted speech. Unfortunately, this method is sensitive to a F_0 ratio of the source and the target speakers. For extreme values of the F_0 ratio, typically in cross-gender conversion, both the overall quality and speech similarity tend to be worse than in VC systems with vocoder used as the synthesis filter.

In the second method, the F_0 transformation is carried out by directly modifying the residual signal, which is derived from the source voice with inverse filtering based on the extracted mel-cepstrum (MCEP). The transformation also includes the time-scaling WSOLA algorithm. Despite several advantages, like the preservation of natural phase components or flexibility in controlling F_0 transformation ratio, this method suffers from common speech quality degradation due to difficulties in extracting the residual signal.

The third method brings back the usage of vocoder to the VC process. However, the vocoder is not used as the synthesis filter at the last step of VC (as speech generator), but as a tool to transform the F_0 and aperiodic components of the source speaker. The risk of the F_0 extraction errors and voiced/unvoiced decision errors makes this method less preferable for intra-gender VC, where minor F_0 differences are easily handled by direct waveform modifications. On the other hand, applying this approach makes VC more consistent in terms of the expected speaker's identity conversion accuracy [7].

As the results of VC 2016 Challenge showed [3], that it is much harder to achieve a robust cross-gender conversion than in

[1] Voice Conversion Challenge website: www.vc-challenge.org

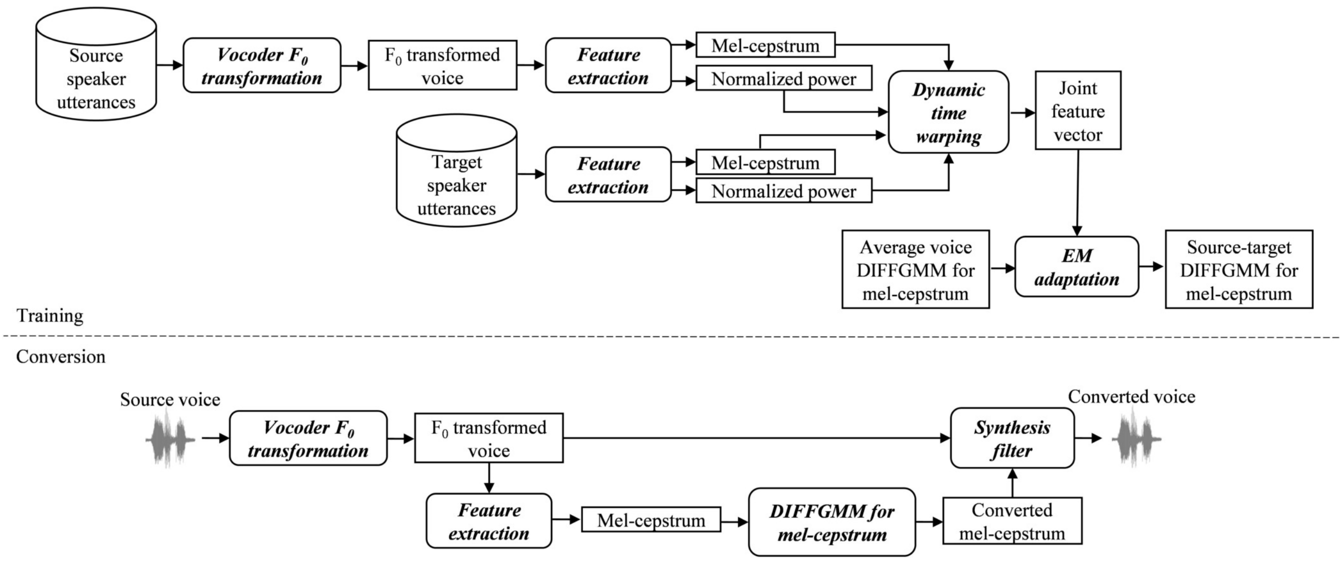

Fig. 1. Block diagram of the proposed system.

intra-gender scenario. Moreover, dealing with both intra- and cross-gender cases using a single system, requires a compromise between the output speech quality and the accuracy of the speaker's identity conversion.

In this paper we present a method for conversion model preparation using average speaker model adaptation and the application of constant rate F_0 transformation. The proposed methods enhance cross-gender conversion quality obtained with DIFFVC framework and WORLD [9] (D4C edition [10]) vocoder. The developed system was submitted for the 2018 Voice Conversion Challenge. The evaluation results confirmed a high performance of the proposed solution, placing it within the top 3 results in the cross-gender task in terms of the measured speech quality and voice similarity.

The paper is organized as follows. In Section II, the proposed algorithm is described. In Section III we present Voice Conversion Challenge dataset. In Section IV, experimental evaluations and results are discussed. Finally, we conclude this paper in Section V.

2. VOICE CONVERSION METHOD DESCRIPTION

The proposed system was designed to handle training of conversion models for parallel data sets (the HUB task in VCC 2018). In this approach each source-target conversion model is trained using the recordings of the source and the target speakers, where both pronounced the same utterances.

By using the constant average F_0 ratio, improving the accuracy of the time-alignment of samples, and introducing average background conversion model, we achieved a significant conversion quality improvement in cross-gender scenarios without degradation of intra-gender scenarios. The proposed solution was developed using a Sprocket VC Toolkit[2], which also served as the 2018 VC Challenge baseline. The toolkit offers several implementations of the traditional GMM, as well as differential-GMM vocoder-free conversion methods. The block diagram of the proposed solution is presented in Fig. 1.

As the DIFFVC conversion framework was chosen, the first step of the whole training process was to prepare the F_0 transformed version of the source voice. The transformation was performed using WORLD vocoder. In this step log(F_0), spectral envelope and aperiodicity (using DC4 algorithm [10]) features were extracted. To minimize the risk of vocoder resynthesis artifacts, only constant F_0 ratio transformation, based on average F_0 values, was introduced in this solution. This decision was motivated by observations of perceptually distracting artifacts, when complete F_0 cross-gender transformations were performed. Introduction of the constant average F_0 ratio helped reducing the observed errors significantly. Next, in order to train the conversion function, MCEP features were extracted from both original target and F_0-converted source signals. As the last step of data processing, speaker-dependent global variance statistics of MCEP coefficients were computed to deal with oversmoothing effects during conversion [7]. This step allows for maintaining high speaker-specificity of the MCEP statistics in contrast to speaker-independent estimate.

To estimate the meaningful joint source-target MCEP distributions, all the respective utterances in the dataset were time-aligned. This task was performed by applying three iterations of the dynamic time warping (DTW) of the MCEP-parametrized speech signals. In the first iteration, the initial alignment was calculated between the target and the corresponding F_0-transformed source samples. Applying a recursive approach, every next alignment was recalculated using converted speech samples of the source voice, rather than the original F_0-transformed versions. To generate those converted speech signals the whole training and conversion process had to be fulfilled before entering next iteration of the DTW algorithm. The proposed source-target alignment strategy resulted in a significant improvement of the joint source-target data alignment for the final conversion model estimation.

Following all the steps described above, the background average-speaker differential MCEP conversion model (DIFFGMM) with 64 GMM components was prepared from the merged databases of all 16 pairs of speakers available in 2018 VCC dataset using the expectation-maximization (EM) algorithm. Introduction of the

[2] Sprocket project repository: http://github.com/k2kobayashi/sprocket

background average model reduces the overfitting of the final conversion model by incorporating knowledge from all speakers and pre-conditioning the final conversion model estimation. The final DIFFGMM conversion model for a specific source-target pair of speakers was eventually achieved by using the EM adaptation of the background average model with the respective source and target data. Although, the applied EM training optimized ML objective, the usage of the pre-trained background model allowed the final 64-component DIFFGMM to converge without overfitting. It was not achievable when no background model was used for initialization for available amount of training data and expected model size. The final DIFFGMM model apparently converged to a local ML solution given the specific source-target adaptation data and the averaged speaker conversion initial conditions. Maximum a-posteriori adaptation can also be used within this framework, which is our objective for further research. The meta-parameters of the models and training were adjusted to improve the observed conversion quality during system development. The maximum possible number of EM iterations in adaptation phase was 150, and the convergence stopping criterion was 0.001.

Once the specific DIFFGMM voice conversion model is ready, a new converted MCEP features can be generated for the test source sample using maximum-likelihood parameter generation (MLPG) approach with static and delta components [7] as presented in the lower part of Fig. 1. To alleviate the oversmoothing effect, MCEP postfiltering with differential GV statistics is applied, as described in [7]. The output waveform is finally obtained by filtering the F_0-transformed source speech with the differential MCEP representation obtained with MLPG using source-target-specific DIFFGMM.

3. VCC DATASET

VCC 2018 dataset for the HUB task consisted of four source and four target voices. Both source and target datasets consisted of two female and two male voices. Each voice was represented by the same set of 81 sentences. The waveforms were stored in a 16-bit WAVE format with the sampling rate of 22.05 kHz. The F_0 ratio between source and target speakers varied from 0.49 up to 2.18, depending on particular conversion pair. The goal of the task was to prepare all 16 possible conversion models between the source and the target voices.

4. RESULTS AND DISCUSSION

The evaluation of system performance was conducted by the organizers of VCC 2018 and the results are depicted below. Figure 2 describes the overall results for all VCC 2018 contributions and all conversion pairs, as well as reference scores of a baseline system (B01), source (S00) and target (T00) voices.

Two distinct measures were used for the evaluation: the quality of the generated audio and the similarity of the perceived identity of the target and converted speech samples. Mean Opinion Score (MOS) was used for speech quality assessment. The similarity of voices was defined as the percentage of sample pairs indicated as 'Same, absolutely sure' or 'Same, not sure'.

Out of 23 participating teams, N08 Techmo achieved the third-best result in both evaluation categories, with the average MOS of 3.35 and the similarity score of 77% for all conversion pairs. That means a 5% improvement of the similarity score and a decrease of 0.22 of the MOS score compared to the VCC 2018 baseline system (B01). It is worth mentioning that in fact the baseline system consisted of two separate sub-systems: vocoder-free and vocoder-based VC setup, which have been developed for the intra-gender and cross-gender conversion pairs respectively [11]. As mentioned before, the main goal of the research was to improve the cross-gender conversion, which was the weakest link of most of other conversion systems so far. Because of that, a comparison of the overall results for all conversion pairs is not reliable. Having split the results to intra- and cross-gender groups, one can observe meaningful improvement in both quality and similarity scores for cross-gender conversions.

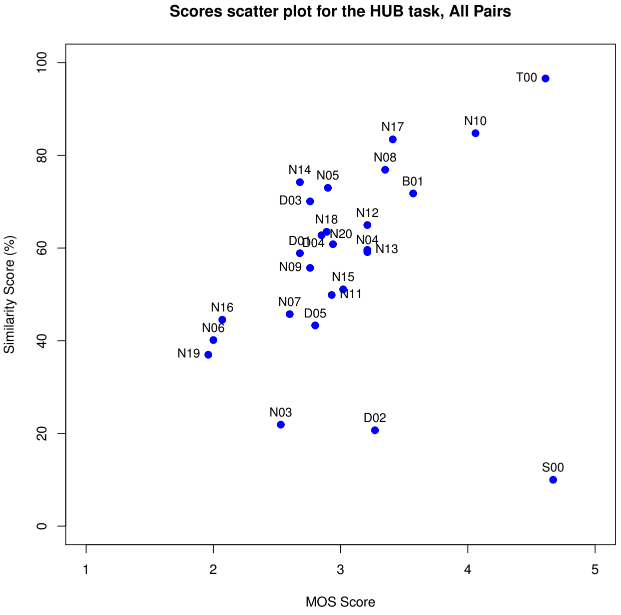

Fig. 2. Overall VCC 2018 results for all teams and conversion pairs. B01 - baseline, N08 - Techmo proposed system.

Figure 3 shows a box plot of MOS results for cross-gender conversions for all VCC 2018 teams. Our system achieved the second highest result in this collation. A more detailed comparison of the proposed system and the VCC 2018 baseline is presented in Table 1. As the evaluation results indicate - combining a constant F_0 ratio transformation with spectral differential filtering reduces the overall degradation to occasional artifacts only, which makes that technique more preferable than techniques with WSOLA F_0 transformation or traditional fully vocoder-based systems for the conversions with extreme F_0 source-target ratios.

When it comes to the speaker similarity, the proposed system achieved a very good performance, giving it the third position in the overall VCC 2018 classification (see Figure 2) and the fourth position in the cross-gender category (with the second highest percentage of top marks, denoting that both converted and target samples are perceived as if they were spoken by the same speaker, see Figure 4). Detailed results are described in Table 2. The similarity score was 77% and 76% for intra- and cross-gender conversion types respectively. It can be therefore said that the proposed system's ability to convert perceived identity of a speaker is independent of the F_0 ratio between the source and the target.

During the research, we found out that training the conversion model for a specific speakers pair, using extended, pre-trained average model with increased number of Gaussian components, helps to improve the overall quality of the conversion process. Comparing

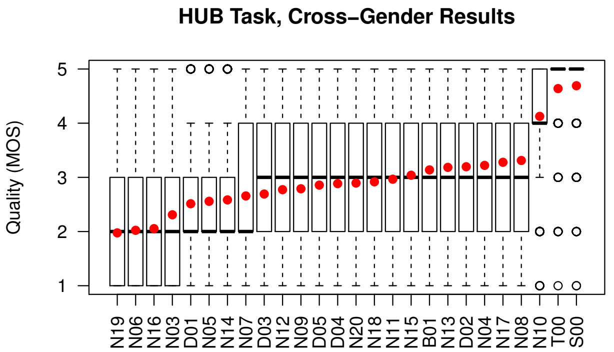

Fig. 3. MOS quality of cross-gender conversions for all VCC 2018 teams. B01 - baseline, N08 - Techmo proposed system.

Table 1. Voice conversion quality (MOS)

All trials MOS				
System	**Average**	**StdDev**	**Count**	**Median**
Source (S00)	4,67	0,67	1120	5
Target (T00)	4,61	0,76	544	5
Baseline (B01)	3,57	1,17	2240	4
Techmo (N08)	3,35	1,16	2240	4
Cross-gender MOS				
System	**Average**	**StdDev**	**Count**	**Median**
Baseline (B01)	3,17	1,17	1120	3
Techmo (N08)	3,22	1,15	1120	3
Intra-gender MOS				
System	**Average**	**StdDev**	**Count**	**Median**
Baseline (B01)	3,95	1,03	1120	4
Techmo (N08)	3,50	1,13	1120	4

similarity scores of the proposed system and the VCC 2018 baseline, an 8% gain in cross-gender conversions can be remarked, without loosing any point in intra-gender conversions.

5. CONCLUSIONS

In this paper we have presented our contribution to the 2018 Voice Conversion Challenge. A new conversion model preparation scheme for DIFFVC framework was proposed and described. The solution is suitable for the use of parallel source-target training data. The proposed method involves multiple time-alignment refining of source-target speech pairs, and usage of the pre-trained average background model for the following specific source-target model training. Constant F_0 ratio source signal transformation was also proposed as a solution to cross-gender F_0 conversion problems.

The proposed conversion scheme was evaluated during the 2018 Voice Conversion Challenge. The obtained results confirm high performance of the proposed solution especially for cross-gender conversion tasks, without any significant intra-gender conversion drawbacks.

Future research might focus on the investigation of the maximum a-posteriori adaptation of the average background conversion model with specific speaker-pair data for both parallel and non-parallel training scenarios, especially when little source or target data is available.

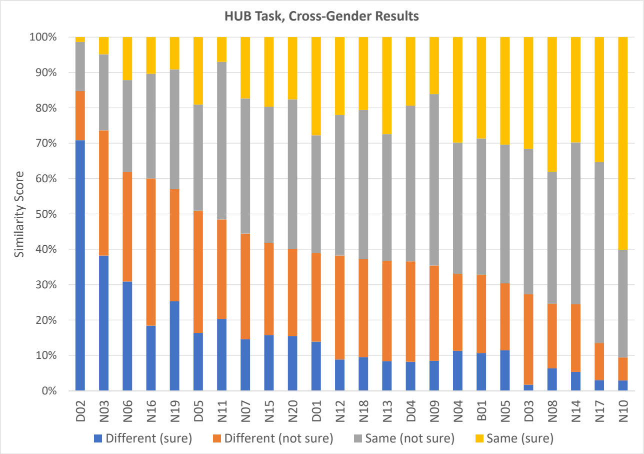

Fig. 4. Similarity score of cross-gender conversions of all VCC 2018 submissions. B01 - baseline, N08 - Techmo proposed system.

Table 2. Similarity scores

All trials	**Different**		**Same**		**Sco-**
System	sure	not sure	not s.	sure	**re**
Source(S00)	75%	15%	6%	4%	10%
Target(T00)	0%	3%	5%	92%	97%
Baseline(B01)	11%	18%	33%	39%	72%
Techmo(N08)	8%	16%	36%	41%	77%
Cross-gender	**Different**		**Same**		**Sco-**
System	sure	not sure	not s.	sure	**re**
Baseline(B01)	11%	21%	34%	33%	68%
Techmo(N08)	8%	15%	40%	37%	76%
Intra-gender	**Different**		**Same**		**Sco-**
System	sure	not sure	not s.	sure	**re**
Baseline(B01)	10%	14%	31%	45%	76%
Techmo(N08)	7%	16%	32%	45%	77%

6. ACKNOWLEDGEMENTS

The authors would like to thank the organizers of 2018 VC Challenge for providing the evaluation results and Figures 2, 3, and 4 . This work was supported by NCBR as part of 'VR Audio Technologies' POIR.01.02.00-00-0104/16 project (Author 1 and 3, Sections 1-3). This research received financial support from the Foundation for Polish Science under grant number First TEAM/2017-3/23 which is co-financed by the European Union under the European Regional Development Fund (Author 2, Sections 4-5).

7. REFERENCES

[1] T. Toda, "Augmented speech production based on real-time statistical voice conversion," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2014, pp. 592–596.

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, Apr 1988, pp. 655–658 vol.1.

[3] Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi, *Analysis of the Voice Conversion Challenge 2016 Evaluation Results*, pp. 1637–1641, Interspeech. International Speech Communication Association, 9 2016.

[4] 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, Nov 2007.

[5] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[6] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveign, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187 – 207, 1999.

[7] K. Kobayashi, T. Toda, and S. Nakamura, "F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2016, pp. 693–700.

[8] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1993, vol. 2, pp. 554–557 vol.2.

[9] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[10] Masanori Morise, "D4c, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57 – 65, 2016.

[11] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The Voice Conversion Challenge 2018: Promoting Development of Parallel and Non-parallel Methods," *ArXiv e-prints*, Apr. 2018.