The Speaker and Language Recognition Workshop (Odyssey 2018)
26-29 June 2018, Les Sables d'Olonne, France

Convolutional Neural Network Based Speaker De-Identification

Fahimeh Bahmaninezhad, Chunlei Zhang, John H.L. Hansen

Center for Robust Speech Systems (CRSS),
The University of Texas at Dallas, Richardson, TX 75080, U.S.A.
{fahimeh.bahmaninezhad, chunlei.zhang, john.hansen}@utdallas.edu

Abstract

Concealing speaker identity in speech signals refers to the task of speaker de-identification, which helps protect the privacy of a speaker. Although, both linguistic and paralinguistic features reveal personal information of a speaker and they both need to be addressed, in this study we only focus on speaker voice characteristics. In other words, our goal is to move away from the source speaker identity while preserving naturalness and quality. The proposed speaker de-identification system maps voice of a given speaker to an average (or gender-dependent average) voice; the mapping is modeled by a new convolutional neural network (CNN) encoder-decoder architecture. Here, the transformation of both spectral and excitation features are studied. The voice conversion challenge 2016 (VCC-2016) database is used to train the system and examine performance of the proposed method. We use two different approaches for evaluations: (1) objective evaluation: equal error rates (EERs) calculated by an i-vector/PLDA speaker recognition system range between 1.265 - 3.46 % on average for all developed systems, and (2) subjective evaluation: achieved 2.8 naturalness mean opinion score (MOS). Both objective and subjective experiments confirm the effectiveness of our proposed de-identification method.

1. Introduction

Speaker de-identification is the task of concealing speaker identity, which may be revealed in linguistic (content of speaker's speech) [1, 2] and paralinguistic (spectral and excitation features of the speech signal uttered by the speaker) [3, 4, 5] features. In this study, we focus on the latter one. Our goal is to map voice characteristics of a given speaker to a new identity, while preserving the naturalness and intelligibility. Speaker de-identification has many applications; for instance, protecting privacy of subjects speaking in a recording (e.g, witness or victim in courtroom/legal scenarios, voices played in some radio or television programs, and medical records), secure transmission of speech data (e.g., hiding speaker identity while transmission of speech data gathered from online banking services), prevention of unauthorized access, data augmentation, etc.

Previous studies in this research area are very limited. In [3] the authors proposed a method for protecting the privacy of speakers by adding masking sound based on white noise (considering different SNRs) or adding noise using band-pass filters. The authors showed that overall intelligibility decreases as the accuracy of protecting privacy increases. Generally, noise masking speech signals can unintentionally degrade intelligibility. In addition, masking just may decrease performance of speaker recognition systems, but subjectively listeners may still recognize the identity of a speaker. On the other hand, the authors in [5] used GMM-based and phonetic based speaker recognition systems for their evaluations. They transformed a source voice to a synthetic target voice called kal-diphone. Using synthetic voice as the target data degrades the performance of the de-identification system. In addition, authors in [6] manually defined piece-wise linear functions to transform the spectral parameters and achieved 4.4% - 98.6% accuracy with different settings; and no subjective test has been reported. Authors in [7] adopted an available transformation method, i.e., the weighted frequency warping. They proposed a new method for selection of a speaker from the database. Transformation applies on the source speaker toward this selected speaker. The selection method is designed to meet three different criteria to achieve an overall promising performance.

Here, in contrast to other related works (to the best of our knowledge, they all used an already available voice transformation method), we propose a new convolution encoder-decoder based voice mapping and incorporated that into our speaker de-identification system. We use the publicly available database of voice conversion challenge 2016 (VCC-2016) [8, 9] to develop our voice mapping system. The convolution neural network (CNN) voice mapping architecture is specifically designed to consider details of the database and has the ability to suppress the errors and noises that might occur during the preparation of data for the voice mapping step. Finally, the speaker de-identification system employs the voice mapping module to transform the voice characteristics of a given speaker to all target speakers in the database. Average or gender-dependent average of mapped voices leads to the de-identified voice.

As a brief summary, the main contribution of this study is the development of a new voice mapping system using convolutional encoder-decoder neural networks. Other key aspect of this study is that we evaluate the proposed architecture with an i-vector/probabilistic linear discriminant analysis (PLDA) [10] speaker recognizer. In addition, the de-identification approach proposed here is designed to mislead both human listeners and machines while preserving quality and naturalness.

Sec.2 introduces the database and feature sets we used in this study. Next, Sec. 3, presents our proposed speaker de-identification architecture. The experimental setup, objective and subjective tests are presented in Sec.4, and finally Sec. 5 concludes the paper.

2. Database and Features

2.1. Database

We use publicly available database of voice conversion challenge 2016 (VCC-2016) [8, 9] here. This database is specif-

This project was funded by AFRL under contract FA8750-15-1-0205 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

10.21437/Odyssey.2018-36

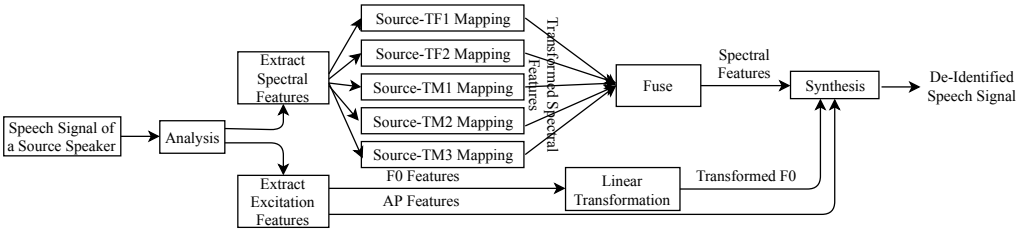


Figure 1: The overall block-diagram of proposed speaker de-identification.

ically designed for the voice conversion application. In voice conversion, we map a source speaker to a target speaker. This database contains speech data of 10 speakers, 5 source speakers (SF1, SF2, SF3, SM1, SM2) and 5 target speakers (TF1, TF2, TM1, TM2, TM3). S and T represent source and target speakers respectively; in addition, M and F refer to male and female. This database is parallel; i.e., all speakers read the same set of sentences. Each speaker has 162 training and 54 test utterances. For developing our systems, we use 150 training utterances for modeling, and the remaining 12 utterances as development data.

The key points that lead us to choose this database include: (1) to preserve linguistic information we need a parallel database [8] (however, there are also growing studies on non-parallel data [11] as well). (2) to the best of our knowledge, there is not any other publicly available parallel database designed specifically for speech synthesis or voice conversion rather than CMU-ARCTIC [12] (which only has 7 speakers); therefore, we chose VCC-2016 as it has more speakers.

2.2. Features

In the proposed speaker de-identification system, we first compress speech into a set of acoustic features, we then de-identify speaker information in the extracted feature space, and finally we synthesize de-identified speech from the acoustic feature space. The STRAIGHT vocoder [13] is used for analysis and synthesis of utterances. STRAIGHT is a high quality vocoder that introduces around 0.5 MOS degradation in the naturalness of the speech signal [14]. STRAIGHT extracts 513-D spectral envelope (SP), 513-D aperiodicity (AP) features as well as 1-D fundamental frequency (F0). We employ speech signal processing toolkit (SPTK) to convert SP to 40-D Mel-cepstral coefficients (MCEP) [15]. The de-identification is only applied to MCEP and F0 features; the AP features are directly mapped from the source speaker to the de-identified speaker.

3. Proposed Speaker De-identification

The details of the proposed speaker de-identification architecture are described in this section. We first introduce the main idea and the overall architecture of the proposed system, and then explain each individual subsystem in detail.

Figure 1 shows the overall block-diagram of the proposed system. Based on this figure, the proposed system performs the speaker de-identification in the following 4 steps:

1) Feature analysis: As we described in Subsection 2.2, spectral (MCEP) and excitation (AP, Log-F0) features are extracted using the STRAIGHT vocoder.

2) Feature mapping: The VCC-2016 database has 5 source and 5 target speakers. 25 (i.e., every potential mapping from any source to any target) mapping functions from MCEP features of the source to the MCEP features of all target speakers are trained based on a new convolutional encoder-decoder neural network architecture (which is described in detail in Subsection 3.1). To preserve the variance of training data and partially resolve the over-smoothing problem, we simply scale the variance of the generated MCEP features to that of the same speaker in the training data. For Log-F0 a simple linear transformation is applied. In addition, AP is moved directly from the source speaker to the de-identified speaker without any modification.

3) Fusion: For a given source speaker, we map the MCEP and Log-F0 features to all target speakers in the database (based on the technique explained in the previous step). Next, mapped features are fused together with two different approaches: (i) average, and (ii) gender-dependent average. It is clear that we can also apply weighted averaging to obtain different voices for each source speaker, but in this study for the sake of simplicity we use equally weighted averaging.

4) Synthesis: In this step, transformed MCEP features are converted back to SP using SPTK toolkit. We stack the SP, F0 as well as the AP (obtained from the previous steps) and use STRAIGHT synthesis module to generate the de-identified speech samples.

3.1. Convolutional Encoder-Decoder Mapping

This subsection introduces a new neural network architecture for mapping acoustic features from a source speaker to a target speaker. Similar to all neural networks, our mapping network has train and test (de-identification) phases.

In the training phase, source and target utterances are first aligned using the dynamic time warping (DTW) algorithm. Next, we prepare the data for our training procedure. The input and output of the network are stacks of 15 consecutive frames of MCEP features. We can interpret these 15 frames as one frame that is appended with 7 previous and 7 next frames. Finally, the mapping network is trained to model the non-linear transformation from the input sequence to the output sequence.

In the de-identification phase, we slide a 15-frame-length window over the input sequence and feed each window as the input to the trained network. The network transforms the input into the same dimensional output; however, we only keep the middle one.

In this paper, we introduce a new convolutional neural network (CNN)-based structure to perform the spectral mapping. CNNs represent a variation of neural networks [16] which have a unique structure with a cascade of convolution and pooling layers. Three key CNN aspects benefit our task: local connectivity, weight sharing, and pooling [17]. Local connectivity makes the system more robust to noise. In addition, while static features are sufficient for the network, the benefits of using dynamic features are captured by CNN filtering. Also, weight sharing reduces the number of parameters which par-

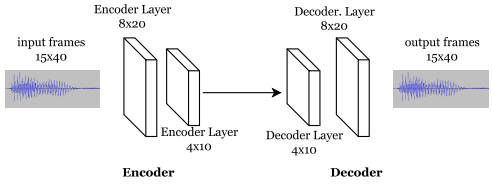


Figure 2: Convolutional encoder-decoder architecture.

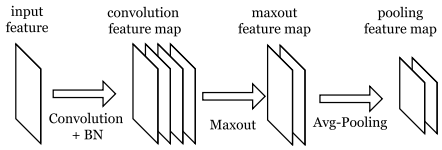


Figure 3: Encoding layer: encodes input into a lower dimensional representation. BN is batch-normalization. Each convolution layer uses maxout and is followed by average pooling.

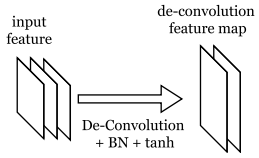


Figure 4: Decoding layer: decodes input. The activation function is tanh, and BN is batch normalization.

tially addresses the issue of over-fitting. Specially here, the VCC-2016 database is small (this can be valid for most of paralleled databases) and can cause an over-fitting problem. Pooling as well can help suppress potential errors of dynamic time warping (DTW) for aligning the two feature sets.

Various approaches have been introduced to convert spectral features. Examples include, joint density Gaussian mixture model (JDGMM) [18] with parameter generation algorithm [19] (to incorporate dynamic features) which are traditional methods for converting spectral features. LSTM-RNN [20], stacked joint-autoencoder [21], generative training of DNN [22], exemplar-based conversion [23] are among more recent trends in voice conversion. In addition, [24] proposed (combining different techniques including) applying direct waveform modification using spectral differential filtering (DIFFVC) with GMM-based VC and ranked one of the top systems in the VCC-2016 [9].

CNN-based mapping has multiple advantages over other voice mapping methods which include: (1) Compared to approaches that use delta features to capture time-dependencies (such as, JDGMM), our convolutional encoder-decoder network is able to automatically capture the dependencies between adjacent acoustic feature frames without including dynamic features. As a result, our method does not need a parameter generation algorithm which is prone to the over-smooting problem. (2) Compared to LSTM-RNN approaches, our network is faster and easier to train. Additionally, due to the recurrent nature of LSTM-RNN, it cannot fully exploit the GPU capabilities.

Figure 2 shows the overall structure of the proposed convolutional encoder-decoder. As it is shown in the figure, the structure contains an encoder followed by a decoder. Encoder is a stack of convolution and pooling layers (Figure 3) and decoder is a stack of convolution-transpose[1] (Figure 4) layers. The convolution and pooling layers encode the input into low resolution representations and convolution-transpose layers upsample the data to its original high resolution space. Applying convolution-transpose after the convolutional layers has shown to be effective in other applications; such as image segmentation [26], emotion recognition [25], etc.

4. Experiments

4.1. Evaluation Metrics

In this subsection, the metrics employed for evaluation of the proposed speaker de-identification system are explained. Experiments are categorized in objective and subjective tests.

For objective evaluations, we developed an i-vector/PLDA based speaker recognition [10] system which is explained in 4.1.1. Similar to other speaker recognition evaluations, we report equal error rate (EER) to evaluate and compare the performance of the developed systems [27]. EER measures the error rate of a system at the threshold that miss alarm and false alarm are equal [27].

For subjective evaluations, we conducted an informal subjective test. Details of the experiment are described in 4.1.2.

4.1.1. Speaker Recognition Evaluation

Speaker recognition is the task of recognizing whether a given utterance belongs to a target speaker or not. Here we employ an i-vector/PLDA speaker recognition solution.

In typical i-vector/PLDA speaker recognition systems, Mel-frequency cepstral coefficients (MFCCs) are first extracted as input feature vectors, then voice activity detection (VAD) is applied to remove non-speech segments. Next, a UBM and total variability matrix (TV) are trained, and i-vectors are extracted. Thereafter, i-vectors are post-processed with length-normalization and LDA. Eventually, PLDA is trained and final log-likelihood scores are calculated [28].

In detail, the speaker- and channel-dependent GMM supervector in the i-vector configuration is factorized as [10],

$$M = m + Tw, \tag{1}$$

where m is the speaker and channel-independent UBM supervector, T is total variability (TV) matrix that maps the high-dimensional GMM supervector to a lower-dimensional vector w; or so-called i-vector representation [10].

The expectation maximization (EM) algorithm is used to train both UBM and TV matrix. In the E-step, w is considered a latent variable with a normal prior distribution $N(0, I)$. At the end of the optimization, the estimated value for each i-vector is the mean of the posterior distribution of w [10]. The estimated i-vector is:

$$\hat{w}(u) = (I + T^T \Sigma^{-1} N(u) T)^{-1} T^T \Sigma^{-1} S(u), \tag{2}$$

where Σ is the UBM covariance matrix. In addition, $N(u)$ and $S(u)$ are zeroth and centralized first order Baum-Welch statistics for utterance u, respectively.

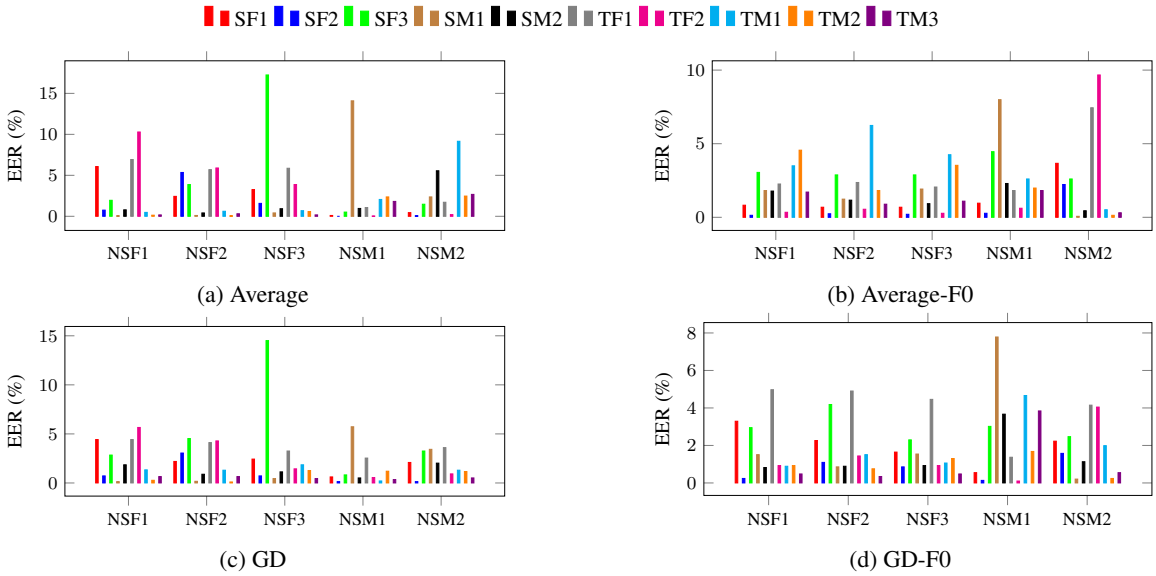
4.1.2. Naturalness Evaluation

For subjective evaluation, we conducted mean opinion score (MOS)-naturalness test. 20 listeners participated in the evalua-

[1] Also known as de-convolution, up-convolution, backward strided convolution and fractionally strided convolution [25]

Table 1: EER (%) for original source and target speakers.

	SF1	SF2	SF3	SM1	SM2	TF1	TF2	TM1	TM2	TM3
EER (%)	2.516	0.559	0.4892	0.1747	0.7687	1.747	0.3494	0.4542	0.2096	0.2795



(a) Average (b) Average-F0

(c) GD (d) GD-F0

Figure 5: EER(%) results for four different systems: a) Average, b) Average-F0, c) GD, d) GD-F0. For every newly generated speaker, the equal error rate against available 10 speakers in database is reported.

Table 2: Summary of results. The EER(%) in figure 5 are averaged here for each newly generated speaker.

Voice De-ID	NSF1	NSF2	NSF3	NSM1	NSM2
Average	2.764	2.476	3.46	2.301	2.613
Average-F0	1.999	1.807	1.783	2.483	2.709
GD	2.232	2.129	2.751	1.265	1.845
GD-F0	1.701	1.824	1.550	2.682	1.859

tions. Listeners are asked to rank the naturalness of 50 randomly chosen utterances from 1 (bad) to 5 (excellent).

4.2. Experimental Conditions

This section describes details on the database used for training the i-vector/PLDA speaker recognition system and CNN configuration we adopted in developing of our system.

For the speaker recognition system, we first extract 19 MFCC features and append them with energy, delta, and delta-delta features using a 25-ms window with sequential 10-ms frame shifts. Next, energy-based VAD is used to remove non-speech segments. A 2048-mixture full covariance UBM and total variability matrix are trained using data collected from SRE2004, 2005, 2006, 2008 and Switchboard II phase 2,3 and Switchboard Cellular Part1 and Part2 [29, 30]. For training both LDA and PLDA, we use training data from VCC-2016 database. The enrollment/test data also includes test utterances from VCC-2016 database.

The CNN encoder-decoder introduced in Section 3.1 uses 2 convolution and 2 convolution-transpose layers. The first convolution layer converts 15x40-D to 15x40x256, which reduces to 15x40x128 with maxout. Next, average pooling is used to reduce the dimensions to 8x20x128. In the second convolution layer, the 8x20x128 input data is converted to 4x10x512 and again reduces to 4x10x256 with maxout and average pooling. The filter size in CNN for the first and second CNN layers are 9x9 and 3x3. In the decoding layers (convolution-transpose layers) tanh activation function is applied in both layers. The filter size for the first and second convolution-transpose layers are 3x3 and 9x9, respectively. In all CNN and convolution-transpose layers, we used batch normalization.

The minimum mean square error (MSE) has been chosen as the optimization criterion and both L1 and L2 regularization are used here to solve the over-fitting problem. The learning rate starts with 0.01 in initial epochs and decreases gradually. The maximum number of epochs is set to 1000. Adam optimization is also used here for training the model.

4.3. Experimental Results

4.3.1. Objective Test

In this subsection, we evaluate the proposed architecture in terms of equal error rate (EER).

First, we evaluate the performance of the i-vector/PLDA speaker recognition for VCC-2016 database. The EER for each individual speaker is shown in Table 1. The results show the average EER for all speakers is 0.75% which is reasonable.

Next, for each source speaker (SF1, SF2, SF3, SM1, SM2) we generate a new speaker (NSF1, NSF2, NSF3, NSM1, NSM2). For example, NSF1 is created by using the AP and F0 (or linear transformation of F0) of SF1 and average/weighted-average of MCEP features generated by voice mapping systems; specifically, transformation from SF1 to all target speakers (TF1, TF2, TM1, TM2, TM3). We claim that NSF1 is different from all available speakers in the database (all source

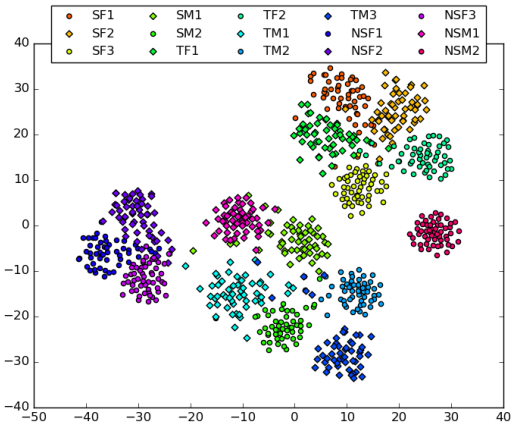


Figure 6: 600-D i-vectors mapped to 2-D representation.

and target ones). The results are presented in Figure 5. We designed trials in a way that smaller EERs represent better de-identification performance.

Figure 5 represents four different approaches for fusing spectral features of different target speakers.
1) **Average**: AP and F0 of the source speaker are directly (without any change) copied to the new speaker; while, the MCEP features are equally weighted average of transformation to all target speakers.
2) **Average-F0**: this is exactly similar to the previous version except that F0 of the new speaker (e.g., NSF1) is a linear transformation of F0 for SF1. Here, if the source speaker is female, we decrease F0 by 10% and if the source speaker is male, we increase F0 by 10%.
3) **Gender-dependent (GD)**: in this system F0 and AP are copied from the source speaker to the de-identified speaker. However, the MCEP features are the average of only the cross-gender voice mapping models. For example, for NSF1, we average MCEP features generated by SF1-TM1, SF1-TM2, SF1-TM3 voice mapping systems (in contrast to average between all SF1-TF1, SF1-TF2, SF1-TM1, SF1-TM2, SF1-TM3 voice mapping systems which we had in the first system; i.e. "Average").
4) **GD-F0**: this is also exactly the same as "GD" however F0 is linearly increased or decreased by 10% for male and female speakers, respectively.

In Figure 5, it is clear that for both "Average" and "GD" when we change F0 we obtain better performance. For example, comparing "Average" and "Average-F0" the EER for NSF3 against SF3 has improved significantly. Therefore, we can conclude that changing F0 even linearly can help. In addition, comparing "Average" and "GD" systems, in all cases, we obtain better performance with "GD" (except, comparing EERs of NSF3 and SF3 that there is not significant improvement). In "GD", we only use cross-gender models; therefore we expect that, as target speakers are more different from the source speaker, the new speaker will be more distinct. The "GD-F0" approximately outperforms the other three systems. Table 2 summarizes the average EERs captured by each of the four systems for newly created speakers (NSF1, NSF2, NSF3, NSM1, NSM2). These results also confirm that transforming F0 and using gender information help decrease the EER.

Figure 6 uses t-Distributed Stochastic Neighbor Embedding (t-SNE) [31] to map the 600-D i-vector representation of test data into 2-D space. This figure also confirms that the 10 original speakers of the database and 5 new generated speakers (de-identified speakers) are almost distinct.

4.3.2. Subjective Test

For subjective evaluation, we conducted MOS-naturalness test for "GD-F0" speaker de-identification system. We did an informal subjective test at CRSS and obtained 2.8 for pooled utterances of all new generated speakers.

In addition to the MOS-naturalness, an additional subjective test can be designed. We also ask participants "if they can distinguish the new speaker from each available speaker in the database or not". We did an informal subjective test at CRSS and we obtained 100% accuracy for "GD-F0". One of the reasons is that we changed F0, and mapped the source speaker from male to female and vice versa.

5. Conclusion

This paper presented a new solution for the speaker de-identification task. For a given speech signal of a speaker, first, spectral and excitation features are extracted. The spectral features are mapped non-linearly with a novel convolutional encoder-decoder based voice conversion system; and F0 is converted linearly. Transformed features are finally combined together and synthesized to generate the de-identified speech signal. The experiments were carried out on VCC-2016 database and evaluated subjectively and objectively with i-vector/PLDA speaker recognition system. Each source speaker in the database was mapped to a new speaker; for the best proposed system (i.e., "GD-F0") the EER varies between 1.55%-2.682%, and 2.8 was achieved for the subjective MOS-naturalness test. For similarity as well, new speakers were discriminated from the source speaker with 100% accuracy for "GD-F0" speaker de-identification system.

6. References

[1] Sree Hari Krishnan Parthasarathi, Mathew Magimai Doss, Hervé Bourlard, and Daniel Gatica-Perez, "Evaluating the robustness of privacy-sensitive audio features for speech detection in personal audio log scenarios," *IEEE ICASSP*, pp. 4474–4477, 2010.

[2] Sree Hari Krishnan Parthasarathi, Hervé Bourlard, and Daniel Gatica-Perez, "Wordless sounds: robust speaker diarization using privacy-preserving audio representations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 85–98, 2013.

[3] Kei Hashimoto, Junichi Yamagishi, and Isao Echizen, "Privacy-preserving sound to degrade automatic speaker verification performance," *IEEE ICASSP*, pp. 5500–5504, 2016.

[4] Tadej Justin, Vitomir Štruc, Simon Dobrišek, Boštjan Vesnicer, Ivo Ipšić, and France Mihelič, "Speaker de-identification using diphone recognition and speech synthesis," *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 4, pp. 1–7, 2015.

[5] Qin Jin, Arthur R Toth, Tanja Schultz, and Alan W Black, "Speaker de-identification via voice transformation," *IEEE ASRU*, pp. 529–533, 2009.

[6] Carmen Magariños, Paula Lopez-Otero, Laura Docio-Fernandez, Eduardo R Banga, Carmen Garcia-Mateo, and Daniel Erro, "Piecewise linear definition of transformation functions for speaker de-identification," *Sensing, Processing and Learning for Intelligent Machines (SPLINE), 2016 First International Workshop on*, pp. 1–5, 2016.

[7] Miran Pobar and Ivo Ipsic, "Online speaker de-identification using voice transformation," *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, pp. 1264–1267, 2014.

[8] Tomoki Toda, L Chen, Daisuke Saito, Fernando Villavicencio, Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi, "The voice conversion challenge 2016," *ISCA INTERSPEECH*, 2016.

[9] Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi, "Analysis of the voice conversion challenge 2016 evaluation results," *ISCA INTERSPEECH*, 2016.

[10] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[11] Toru Nakashika, Tetsuya Takiguchi, and Yasuhiro Minami, "Non-parallel training in voice conversion using an adaptive restricted boltzmann machine," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2032–2045, 2016.

[12] John Kominek and Alan W Black, "The CMU arctic speech databases," *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[13] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain 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, no. 3, pp. 187–207, 1999.

[14] Hideki Banno, Hiroaki Hata, Masanori Morise, Toru Takahashi, Toshio Irino, and Hideki Kawahara, "Implementation of realtime straight speech manipulation system: Report on its first implementation," *Acoustical science and technology*, vol. 28, no. 3, pp. 140–146, 2007.

[15] Soheil Khorram, Hossein Sameti, Fahimeh Bahmaninezhad, Simon King, and Thomas Drugman, "Context-dependent acoustic modeling based on hidden maximum entropy model for statistical parametric speech synthesis," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 12, 2014.

[16] Yann LeCun, Yoshua Bengio, et al., "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, pp. 1995, 1995.

[17] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[18] Alexander Kain and Michael W Macon, "Spectral voice conversion for text-to-speech synthesis," *IEEE ICASSP*, vol. 1, pp. 285–288, 1998.

[19] Tomoki Toda, Alan W Black, and Keiichi 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.

[20] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," *IEEE ICASSP*, pp. 4869–4873, 2015.

[21] Seyed Hamidreza Mohammadi and Alexander Kain, "A voice conversion mapping function based on a stacked joint-autoencoder," *ISCA INTERSPEECH*, pp. 1647–1651, 2016.

[22] Ling-Hui Chen, Li-Juan Liu, Zhen-Hua Ling, Yuan Jiang, and Li-Rong Dai, "The USTC system for voice conversion challenge 2016: Neural network based approaches for spectrum, aperiodicity and f 0 conversion," *ISCA INTERSPEECH*, pp. 1642–1646, 2016.

[23] Yi-Chiao Wu, Hsin-Te Hwang, Chin-Cheng Hsu, Yu Tsao, and Hsin-Min Wang, "Locally linear embedding for exemplar-based spectral conversion," *ISCA INTERSPEECH*, 2016.

[24] Kazuhiro Kobayashi, Shinnosuke Takamichi, Satoshi Nakamura, and Tomoki Toda, "The NU-NAIST voice conversion system for the voice conversion challenge 2016," *ISCA INTERSPEECH*, pp. 1667–1671, 2016.

[25] Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, Melvin McInnis, and Emily Mower Provost, "Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition," *ISCA INTERSPEECH*, 2017.

[26] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[27] John HL Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[28] Chunlei Zhang, Fahimeh Bahmaninezhad, Shivesh Ranjan, Chengzhu Yu, Navid Shokouhi, and John HL Hansen, "UTD-CRSS systems for 2016 NIST speaker recognition evaluation," *ISCA INTERSPEECH*, pp. 1343–1347, 2017.

[29] Fahimeh Bahmaninezhad and John HL Hansen, "Generalized discriminant analysis (GDA) for improved i-vector based speaker recognition," *ISCA INTERSPEECH*, 2016.

[30] Fahimeh Bahmaninezhad and John HL Hansen, "i-vector/PLDA speaker recognition using support vectors with discriminant analysis," *IEEE ICASSP*, 2017.

[31] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.