Depersonalization of Speech Using Speaker-Specific Transform Based on Long-Term Spectrum

Miroslav RUJZL, Milan SIGMUND

Dept. of Radio Electronics, Brno University of Technology, Technicka 12, 616 00 Brno, Czech Republic

196809@vut.cz, sigmund@vut.cz

Submitted February 12, 2023 / Accepted August 5, 2023 / Online first October 30, 2023

Abstract. *This paper introduces a novel approach for hiding personal information in speech signals. The proposed approach applied a transform warping function, which is obtained from a long-term linear prediction spectrum individually for each speaker. The depersonalized speech was compared with the often used technique based on vocal tract length normalization. The proposed approach performs wider manipulation of fundamental frequency and provides higher intelligibility by 5% in clean speech and by 8% for signal-to-noise ratio 5 dB. It also significantly alters the derived glottal pulses, making them difficult to use for personality analysis. Speech intelligibility index and glottal pulse distortion are new aspects in the field of voice depersonalization.*

Keywords

Speech depersonalization, long-term spectrum, voice transformation, depersonalized speech evaluation

1. Introduction

The term depersonalization refers to the modification of an individual's speech signal in such a way that it is impossible (or very difficult) to determine the identity of the speaker while preserving speech intelligibility. Depersonalized speech can be obtained from natural speech in a process, commonly called de-identification, by which data characterizing the speaker are altered or removed from the speech signal. Our proposed approach uses the speaker-specific voice spectrum to eliminate the characteristics of a person. For this reason, we prefer a more apt term "depersonalization" over the general term "de-identification". Since the depersonalized speech is intended to serve in the same situations as the original speech, the process must produce speech that is well intelligible to the general public and, furthermore, it should sound natural.

A typical example of the use of depersonalized speech can be storage of voice recordings in a public database while preserving the anonymity of the speaker. The request for anonymity often comes from the authorities, especially after 2018 when the General Data Protection Regulation (GDPR) was introduced in the European Union [1]. For example, data protection and restrictions also apply to speech-based user interfaces [2]. Another important applications of depersonalized techniques are in forensic scenarios, typically testimonies of witnesses in a court under witness protection regime. Obviously, these techniques could also be abused in an illegal way. In criminal cases, there is a need for a reversible process, i.e. re-identification of the speaker from the altered/degraded speech signal [3].

The state-of-the-art of speech depersonalization is characterized by very high intelligibility, especially in a quiet environment. However, there is no standard criterion for practical evaluation of depersonalization methods. Many authors measure the intelligibility of their systems using the word error rate (WER), but the WER values obtained may depend on the structure of the words in the test (e.g. phonetic difficulty or acoustic similarity of the words). In addition to the listening tests, we applied an objective intelligibility criterion based only on the quality of the speech signal. Such a criterion is suitable for a more accurate evaluation of depersonalization methods in adverse acoustic conditions. It can also help to find the best among equally successful methods according to WER. Another original approach applied in our research is the investigation of the distortion of derived glottal pulses due to depersonalization. This phenomenon has not yet been described or mentioned in any publication. Knowing that glottal pulses can be used, for example, for psychoanalysis, it will be necessary to pay attention to this part of depersonalization as well - especially in view of the growing capabilities of artificial intelligence. The aim of our research is to develop an algorithm for effective depersonalization of speech signals, taking into account the above facts.

The rest of this paper is organized as follows: Section 2 provides a brief overview of selected publications dealing with various speech depersonalization methods. The proposed approach is described in Sec. 3. Section 4 presents experimental results structured in three sub-sections. The last section, Section 5, summarizes the conclusions and suggests future work for further development of the proposed approach.

DOI: 10.13164/re.2023.0523

2. Brief Overview of Used Metods

In general, current methods used in practice are based on one of the two principles: speech conversion or speech manipulation. Methods using conversion perform twofold transfer, namely speech-to-text followed by text-to-speech, while the internal techniques for converting speech to text and back may be different. A typical representative of speech conversion is Diphone Recognition Step and Speech Synthesis (DROPSY), when diphones are searched for in the segmented recording and converted from text form to voice synthesizer. For example, paper [4] presents DROPSY method, which implements two speech synthesis methods: Hidden Makarov Model (HMM) [5] and Time-Domain Pitch Synchronous Overlap and Add (TD-PSOLA) [6]. With methods based on speech conversion, we cannot speak directly of depersonalization or de-identification, but rather of anonymization, because there is no way of reverse identification with these methods.

Methods using manipulation either degrade the speech signal in some effective way, or try to remove biometric elements associated with the speaker. It is useful to note that these methods do not only include electronic and software methods. Paper [7] also mentions non-electronic methods such as imitation, ventriloquism or speaking with an object inserted in the mouth. However, more sophisticated methods are electronic, which primarily aim to change the frequency features that determine the identity of the speaker from a physiological point of view. These include the fundamental frequency, the position of the third and fourth formants and the closing phase of the glottal wave [8]. One of the comprehensive methods for speech manipulation is Codebook Mapping. Paper [9] describes the procedure of this method, in which a transfer mechanism (codebook) is constructed from the training data using vector and scalar quantization techniques. The vector coefficients obtained from the codebook are used for voice transformation. The advantage of this method is the possibility of using training data to convert the voice to a particular speaker (e.g., the construction of a codebook to convert the voice of speaker A to speaker B [10]). However, the main disadvantage is the need for a large training dataset. There is also a discontinuity problem with this method, which is caused by the segmentation of the recording and its transformation. To eliminate it, the codebook, the so-called Gaussian Mixture Model (GMM) method was proposed, which is based on soft clustering and continuous mapping of spectral features [8]. However, this method also has a discontinuity problem, which can be solved by various methods, e.g. in paper [11] the generation of the maximum likelihood parameter is used to smooth out the discontinuity. With the development of artificial intelligence and deep learning, the possibilities of using these techniques also for voice conversion are increasing. For example, paper [12] shows a comparison between the GMM method and the use of artificial intelligence to extract the spectral components of a given speaker and the subsequent voice transformation.

A very often used method, on which our proposed approach is also based, is Vocal Tract Length Normalization (VTLN). The method uses frequency warping, in which the frequency axis of the power spectrum is locally compressed or expanded according to a given transformation function called the warping function. The shape and parameters of the functions can be of different types [13] and the correct determination of the parameter values and function type is one of the main problems of this method. Some works deal with the estimation of these parameters and appropriate functions by standard methods, e.g. [14], or is possible to use artificial intelligence and machine learning techniques to estimate these parameters [15]. The VTLN method only appropriately transforms the frequency axis and does not interfere with the frequency spectrum in any other way, so the quality and clarity of the transformed signal remains high. Some papers, e.g. [16], also descibe spectrum amplitude adjustments, called amplitude scaling, to compensate for amplitude differences between the speaker's target and source spectra.

3. Proposed Approach

The principle of the proposed approach is shown in Fig. 1. First, the processed speech is segmented into 20 ms frames by a rectangular window. Then, the short-term speech spectrum is estimated frame-by-frame using Fast Fourier Transform. The actual depersonalization is performed in the "Spectrum transformation" block by a specific transform function. The altered speech spectrum is then converted in each frame to the time domain using Inverse Fourier Transform. Finally, a continuous speech waveform is created by concatenating the frames of the depersonalized speech signal. It is a simple connection of consecutive blocks, because experiments have shown that the PSOLA algorithm, which is standardly used in the VTLN method, does not improve the measured parameters, especially the speech intelligibility index, when applied. The speaker-specific part of speech is estimated in advance by long-term spectral analysis using predictive coefficients in linear prediction (LP). The order of the long-term predictive coefficients should be chosen to be low due to computational complexity. In Sec. 4.1, the effect of order on intelligibility is discussed in more detail. The following formulas deal with second order LP-coefficients for simplicity and easy of understanding For each speaker, the first three autocorrelation coefficients are calculated frame-by-frame and averaged across all frames. The average coefficients $\overline{R}(0)$ to $\overline{R}(2)$ then are used to obtain the average LP-coefficients $\overline{a}(1)$ and $\overline{a}(2)$ by solving the linear equations:

$$\begin{pmatrix} \overline{R}(0) & \overline{R}(1) \\ \overline{R}(1) & \overline{R}(0) \end{pmatrix} \begin{pmatrix} \overline{a}(1) \\ \overline{a}(2) \end{pmatrix} = \begin{pmatrix} \overline{R}(1) \\ \overline{R}(2) \end{pmatrix}. \tag{1}$$

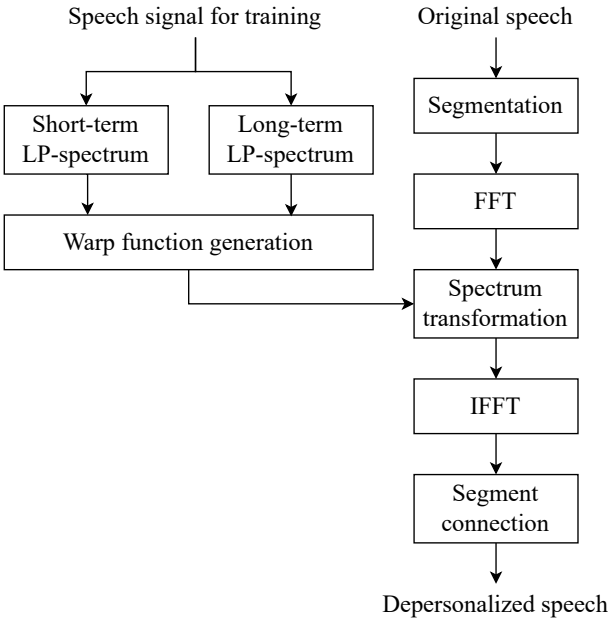

Fig. 1. Flowchart of the proposed approach.

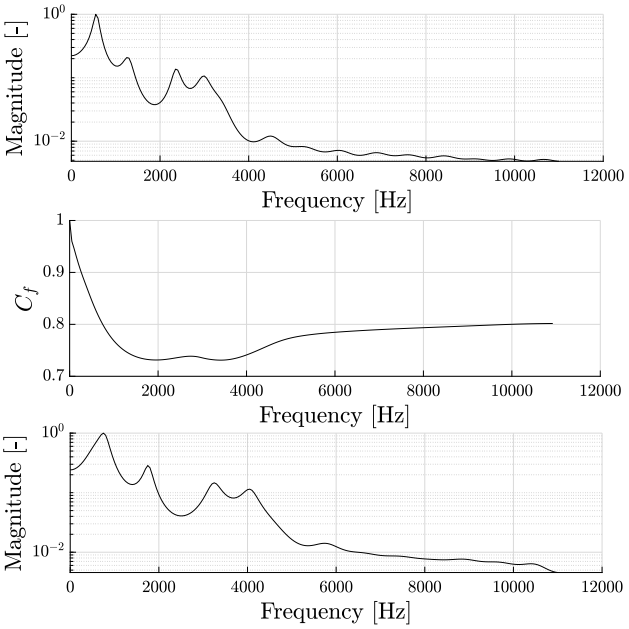

Fig. 2. Short-term original spectrum of vowel /e/ (top), values C_f of the speaker-specific transform vector (middle) and transformed spectrum of the vowel /e/ (bottom).

Finally, the LP-coefficients are applied to estimate the long-term LP-spectrum [17]:

$$\overline{X}(f) = \frac{1}{\left\|1 - \overline{a}(1)\exp\left(-\mathrm{j}2\pi\frac{f}{f_\mathrm{s}}\right) - \overline{a}(2)\exp\left(-\mathrm{j}4\pi\frac{f}{f_\mathrm{s}}\right)\right\|^2} \tag{2}$$

where f ranges from 0 Hz to half the sampling frequency f_s. The long-term spectrum is independent of the just spoken phoneme. It reflects the anatomy of the speaker's vocal tract and can therefore serve as a speaker-specific voice characteristic. This fact was demonstrated in previous experiments where the second-order long-term spectrum was used to identify speakers [18]. For the purposes of transformation in depersonalization, the long-term spectrum is optimized using the LP-spectra $X_k(f)$ of appropriate order which were estimated in all frames. The calculation of these spectra proceeds in the same way as for the long-term spectrum, only the averaging of the autocorrelation coefficients is omitted. The proposed speaker-specific transform function $g(f)$ is

$$g(f) = C_f f = \left(\frac{1}{K}\sum_{k=1}^{K}\sqrt[n]{\frac{X_k(f)}{\overline{X}(f)}}\right) f \tag{3}$$

where C_f denotes elements of the vector of frequency coefficients, K is the number of frames used (in our experiments it was $K \approx 3000$, i.e. speech lasting ca. 1 min) and n is the order of root. An order higher than the simple square root is applied because of the large variability of the proportion of the short-term and long-term LP-spectrum, which has a negative effect on the intelligibility of depersonalized speech. The order should be chosen with regard to computational complexity and the choice of its value is discussed in more detail in Sec. 4.1.

The transform function is expressed as a numeric look-up table. The effect of spectral transformation by the function $g(f)$ on the spectrum of vowel /e/ is depicted in Fig. 2. Middle graph shows the nonlinear transformation trajectory according to which the warping (or scaling) of the frequency axis in the spectrum is carried out. Note that a trajectory value of 1 means no warping. The more the values differ from 1, the more the respective frequency components are shifted in the spectrum. In the transformed spectrum in bottom of Fig. 2, it can be seen that there is a shift of the first four formants towards higher frequencies. However, the shift takes place in the frequency intervals characterizing the vowel /e/, thereby preserving speech intelligibility. On the other hand, changes at frequencies up to 4 kHz make it difficult to determine the speaker's psychological states. For example, some details in the spectral representation of vowels in the 2–3 kHz frequency band are important for stress detection [19].

The functionality of the proposed approach was published in a brief preprint [20] as an initial version using 8th root order and 2nd LP-order. This paper presents an improved approach after correcting the calculation, optimizing the parameter values (root order and LP-order in mutual combination) and extending it with new aspects (noisy speech, emotional speech).

4. Experimental Results

The depersonalized speech signals were objectively evaluated by means of speech intelligibility, fundamental frequency, and other aspects. In experiments, the proposed approach was compared with a previously mentioned method VTLN. In both methods, the transform function is therefore derived from the anatomy of the individual vocal tract.

The transform function created in VTLN depends on two variables, the frequency and the warp factor. The warp factor can be defined in several ways, for example using deep learning techniques. This fact makes the VTLN method relatively complicated. Both methods, VTLN and the proposed approach, were implemented in MATLAB and tested with the same speech database containing 19 Czech native male speakers (marked with the letter M and the registration number). All speech recordings were stored in WAV format at a sampling rate of 22.5 kHz with 16-bit resolution. In our experiments, all the results obtained by the proposed approach were compared with the results of the standard VTLN method.

4.1 Speech Intelligibility

The effect of depersonalization on the intelligibility of processed speech was evaluated using the speech intelligibility index (SII). This is a standard measure of objective speech intelligibility [21] that is based solely on the quality of the speech signal [22]. Therefore, it is not influenced by language or spoken words, allowing experimental validation of the results with recordings from any language. In our experiments, the calculated SII values were expressed on a scale from 0 to 100 (instead of the standard range of 0 to 1) to express intelligibility as a percentage. A possible maximum SII of 100% indicates that all the information contained in the measured speech is available to the listener. As already mentioned, the proposed approach contains two adjustable parameters: the root order and the LP-order (LPO). In order to achieve the best results in terms of intelligibility, i.e. the highest SII value, the optimal values of these parameters were searched by a simple method of parameter sweep. It was not necessary to use a sophisticated optimization method due to the low number of parameters to be considered. Figure 3 shows the average SII calculated from all 19 speakers as a function of two variables: root order and the LP-order. The lowest root order $n = 2$ was not considered because it worsened subjective intelligibility in listening tests. Note that throughout the range the SII values are is over 85%. According to [21], an SII of 75% and above represents good intelligibility, while an SII of 45% and below means poor intelligibility. From this point of view, the used speech material with an average SII of 84% was of high quality. The highest SII value of almost 91% was obtained when using the root order $n = 4$ and the LP-order $LPO = 4$. Thus, all further experiments were performed with this combination of parameter values, i.e. $n = 4$ and $n = 4$.

Table 1 shows the statistical structure of the SII measurement. First, SII values were measured for individual speakers, and then statistics were calculated for the entire group. It can be seen from the comparison of values in Tab. 1 that depersonalization with the proposed approach (as with VTLN) does not significantly reduce speech intelligibility. On the contrary, for 15 of 19 speakers, an increase of 7.1% on average was measured, while only in 4 speakers was a decrease of 0.3% on average.

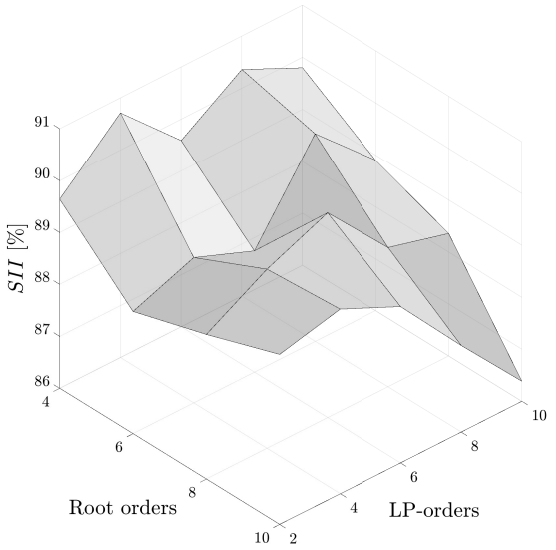

Fig. 3. Speech intelligibility index values as a function of root order and LP-spectrum order.

		Original speech	VTLN	Proposed approach
SII	Mean	84	86	91
	St. dev.	13	9	10
Increase	Number	-	10	15
	Average	-	2.6	7.1
Decrease	Number	-	4	4
	Average	-	0.4	0.3

Tab. 1. Statistical results of intelligibility according SII in percentage obtained from a group of 19 speakers.

4.2 Fundamental Frequency of Voice

The behavior of the proposed approach was also analyzed from the perspective of the fundamental frequency of voice. The fundamental frequency F_0 is one of the most important features of speech, and its values are related to the physiological attributes of a person and his/her vocal expression. The mean fundamental frequency is determined anatomically by the length of the speaker's vocal folds membrane. It is easily audible and therefore directly recognizable by humans as the pitch of the voice that characterizes individual speakers. Instantaneous values of F_0 vary in the range of approx. 100 Hz around the mean value according to the rate of vibration of the vocal cords. However, unvoiced phonemes do not involve vibrations.

The F_0 values were calculated using the residual harmonics method [23] integrated in MATLAB Audio Toolbox and then displayed by a triplet of histograms for each speaker, as shown for example in Fig. 4. The means μ and standard deviations σ of the fundamental frequency were averaged across all speakers and are presented in Tab. 2 for comparison. As can be seen, the depersonalized speech produced by the proposed approach changed F_0 to a greater extent than that by the VTLN. This trend is evident also in Fig. 4. From this observation, it can be concluded that if the proposed approach is applied, it will be more complicated to perform the analysis of F_0 than for the standard VTLN method.

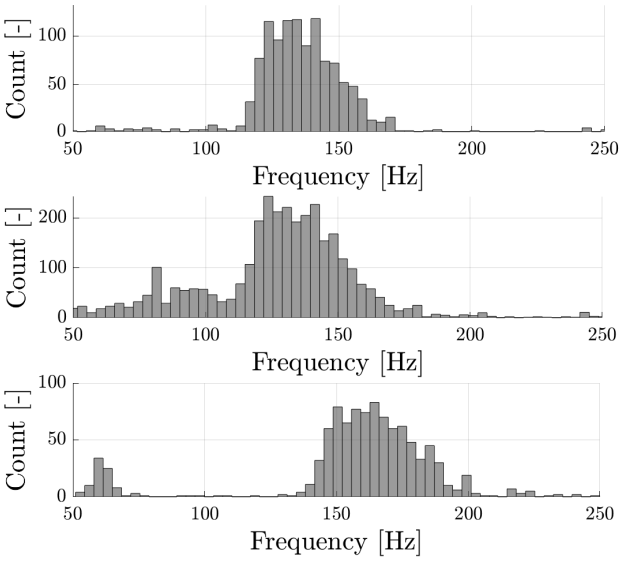

Fig. 4. Histograms of fundamental frequency for speaker M4 obtained from original speech (top), depersonalized speech by VTLN (middle) and depersonalized speech by proposed approach (bottom).

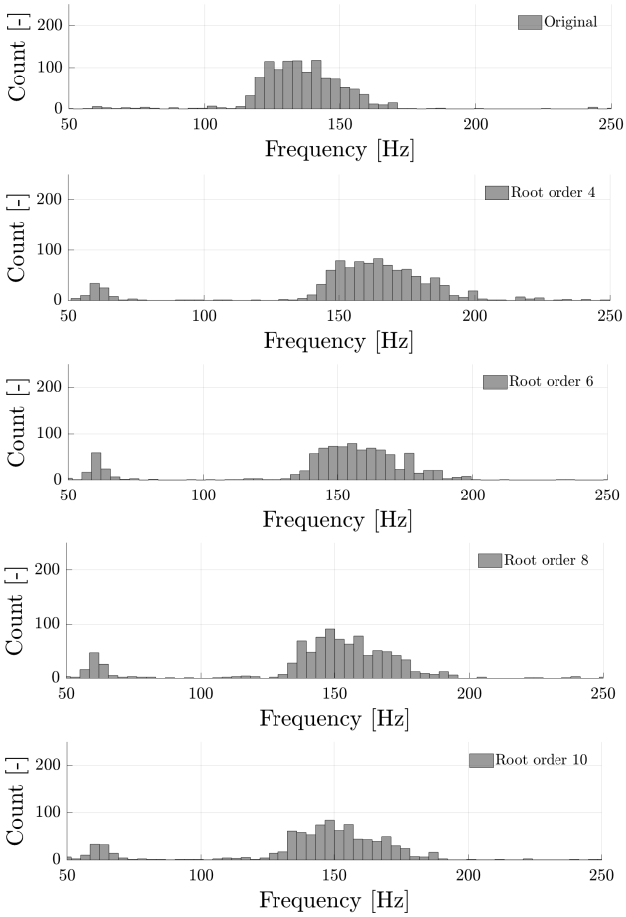

Fig. 5. Histograms of fundamental frequency for speaker M4 obtained from original speech and depersonalized speech by the proposed approach with increasing root order n and fixed order of LP-spectrum $LPO = 4$.

Fund. frequency	Original speech	VTLN	Proposed approach
Average μ	129.9	129.4	174.0
Average σ	21.7	21.2	27.7

Tab. 2. Statistical results of voice fundamental frequency F_0 in Hz obtained from 19 speakers.

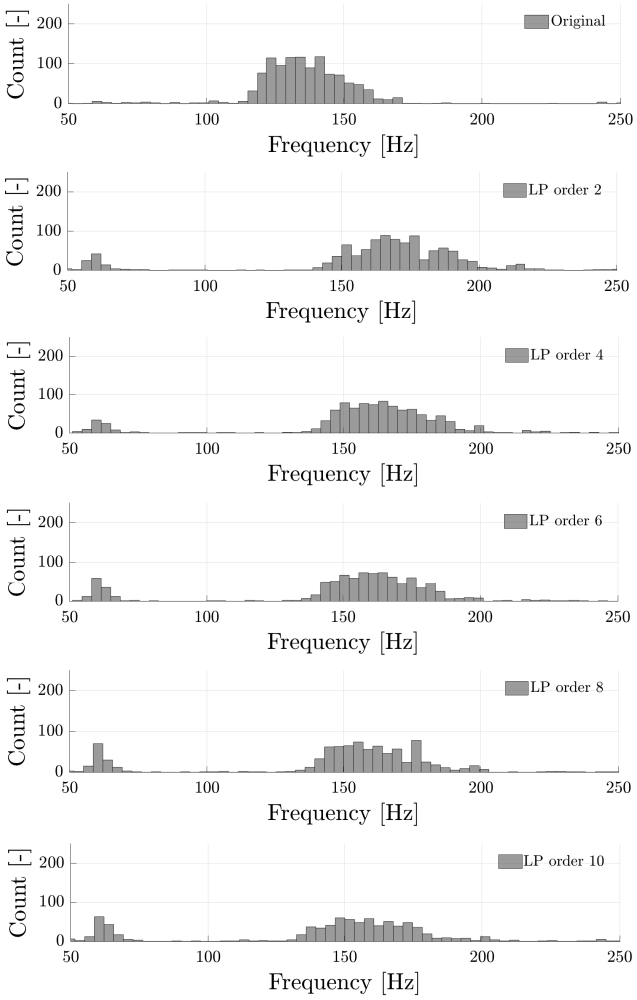

Fig. 6. Histograms of fundamental frequency for speaker M4 obtained from original speech and depersonalized speech by the proposed approach with increasing order of LP-spectrum LPO and fixed root order $n = 4$.

Figures 5 and 6 were generated to show the effect of root order and LP-order on the fundamental frequency histograms. One of the two parameters always remained at a fixed value according to the previous findings in Sec. 4.1 in order to achieve maximum SII, while the other parameter was variable. Both figures illustrate the effect that as the value of the variable parameter increases, the shift values of the fundamental frequencies decrease on average. For the root order sweeping in Fig. 5, this effect corresponds to the assumption that a higher root order n in (3) limits the range of the warping vector C_f, resulting in both less compression and expansion of the speech spectrum.

In Figs. 5 and 6, there are also false low bars in the frequency band from 50 Hz to 70 Hz, which were not in the original speech. It is a negative effect of voiced consonants in depersonalized speech on F_0 calculations by the residual harmonics method [23]. However, all F_0 values less than 80 Hz were ignored in further statistical processing of F_0.

4.3 Other Aspects

In addition to the intelligibility and fundamental frequency, the proposed approach was investigated for robustness against noise, preservation of possible emotions in the speakers, and execution time.

Background noise plays an important role for practical use in an external environment. In the experiments, noisy conditions were created by adding Gaussian noise at three levels of SNR. This type of noise was chosen because the speech signal generally has a Gaussian-like distribution of amplitudes. The impact on intelligibility is shown in Tab. 3. Of course, increasing noise impairs speech intelligibility, but not dramatically. When comparing the original speech and depersonalized speech using both VTLN and the proposed approach, the proposed approach achieved the best results in terms of robustness to noise at all measured SNR levels.

SII	Original speech	VTLN	Proposed approach
Clean speech	84	86	91
SNR = 15 dB	80	86	89
SNR = 10 dB	79	83	86
SNR = 5 dB	72	76	84

Tab. 3. Intelligibility of noisy speech according SII in percentage.

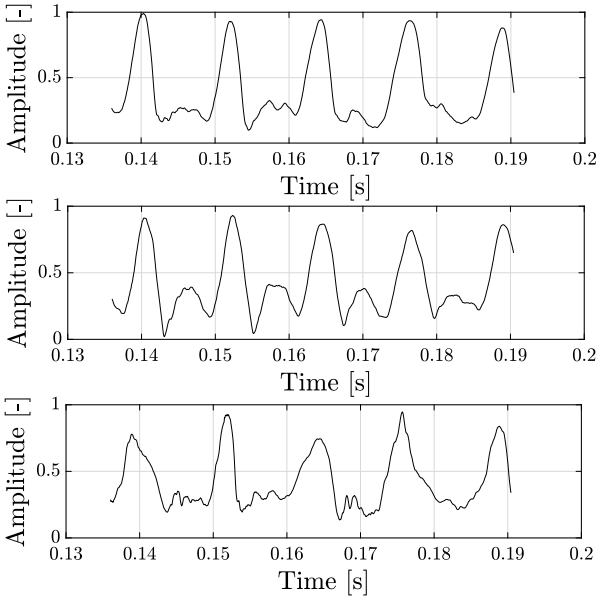

Fig. 7. Glottal pulses of the vowel /a/ derived from original speech (top), depersonalized speech by VTLN (middle) and depersonalized speech by proposed approach (bottom).

Recognizing speaker's emotional state and stress from speech can be important, for example, in forensic analyses. Currently, this aspect also concerns anonymized speech [24]. In natural speech, many emotion identifiers are based on either spectral features or properties of glottal pulses. Spectral methods are not applicable to depersonalized speech because spectral changes after depersonalization are usually greater than changes due to emotion [25]. Therefore, we investigated the distortion of the glottal flow waveform. In our experiments, glottal pulse shapes were derived from the speech signal using the software APARAT [26], which is widely used in speech processing. Figure 7 shows a comparison of glottal pulses obtained from 4 periods of the vowel /a/. As can be seen, VTLN distorts the glottal pulses only slightly, while the proposed approach distorts the shape of the pulses considerably. In this case, for example, the stress in speakers detectable by the opening-to-closing ratio of the pulses [27] cannot be recognized.

The proposed approach is not primarily intended for real-time use. However, for informative purposes, the time consumption of the signal processing was measured in MATLAB. Average calculation speed for a 1 minute record was 4.74 s by the proposed approach, while VTLN did it in 1.87 s. These times should be considered as relative data for mutual comparison. Note that the overall process has not yet been optimized for computational speed. Moreover, the execution time varies depending on the device the process is running on. In particular, the conversion from MATLAB to the implementation on a digital signal processor could significantly speed up the calculations.

5. Conclusion

In this work, we present a new approach for speech depersonalization. Practical advantages can be seen in the transformation function, which is very specific for individual speakers and its easy acquisition. The proposed approach is based on the often used VTLN method and further improves it. Therefore, all experimental results are compared with the results of the standard VTLN using the same speech signals. The proposed approach performs wider F0 manipulation (μ = 174.0, σ = 27.4) than VTLN (μ = 129.4, σ = 21.2) and provides higher intelligibility by 5% in normal conditions (clean speech) and by 8% in adverse conditions (SNR = 5 dB) in terms of SII. According to listening tests, speech depersonalized by both methods was completely intelligible (in normal conditions). In addition, the proposed approach also "depersonalizes" the derived glottal pulses and thus makes it very difficult to perform personality analysis. No one has dealt with this aspect so far. In summary, the proposed approach achieved better results than standard VTLN on a group of 19 speakers in all tested parameters except execution time. Although the main contribution of the presented work lies in the creation of a new efficient depersonalization algorithm, the introduction of two special evaluations

of depersonalized speech, the speech intelligibility index and glottal pulse distortion, is also a novelty in the field.

In future experiments, we will investigate the consistency of the presented results on large groups of male and female speakers across the age range. Furthermore, it will be necessary to test the proposed approach in several languages and consider the evaluation criteria used in other works, see e.g. [28], so that a direct comparison with other effective methods is valuable.

Acknowledgments

This work was supported in part by Quality Internal Grants of BUT (KInG BUT), Reg. No. CZ.02.2.69/0.0/0.0/19_073/0016948.

References

[1] KRZYSZTOFEK, M. *GDPR: General Data Protection Regulation (EU) 2016/679: Post-reform Personal Data Protection in the European Union*. Alphen aan den Rijn (The Netherlands) : Wolters Kluwer, 2019. ISBN: 9789403505947

[2] YOO, I. C., LEE, K., LEEM, S., et al. Speaker anonymization for personal information protection using voice conversion techniques. *IEEE Access*, 2020, vol. 8, p. 198637–198645. DOI: 10.1109/ACCESS.2020.3035416

[3] MAGARINOS, C., LOPEZ-OTERO, P., DOCIO-FERNANDEZ, L., et al. Reversible speaker de-identification using pre-trained transformation functions. *Computer Speech and Language*, 2017, vol. 46, p. 36–52. DOI: 10.1016/j.csl.2017.05.001

[4] JUSTIN, T., STRUC, V., DOBRISEK, S., et al. Speaker de-identification using diphone recognition and speech synthesis. In *Proceedings of the International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Ljubljana (Slovenia), 2015, p. 1–7. DOI: 10.1109/FG.2015.7285021

[5] ZEN, H., NOSE, T., YAMAGISHI, J., et al. The HMM-based speech synthesis system (HTS) version 2.0. In *ISCA Tutorial and Research Workshop on Speech Synthesis (SSW)*. Bonn (Germany), 2007, p. 294–299.

[6] TOMA, S. A., TARSA, G. I., OANCEA, E., et al. A TD-PSOLA based method for speech synthesis and compression. In *8th International Conference on Communications (COMM)*. Bucharest (Romania), 2010, p. 123–126. DOI: 10.1109/ICCOMM.2010.5509044

[7] PERROT, P., AVERSANO, G., CHOLLET, G. Voice disguise and automatic detection: Review and perspectives. Chapter in: STYLIANOU, Y., FAUNDEZ-ZANUY, M., ESPOSITO, A. (eds). *Progress in Nonlinear Speech Processing*. Berlin, Heidelberg: Springer, 2007, p. 101–117. DOI: 10.1007/978-3-540-71505-4_7

[8] GARCIA-MATEO, C., CHOLLET, G. (eds). *Voice Biometrics - Technology, Trust and Security*. London (UK): The Institution of Engineering and Technology, 2021. ISBN: 9781785619007

[9] SATHIAREKHA, K., KUMARESAN, S. A survey on the evolution of various voice conversion techniques. In *3rd International Conference on Advanced Computing and Communication Systems (ICACCS)*. Coimbatore (India), 2016, p. 1–5. DOI: 10.1109/ICACCS.2016.7586373

[10] ABE, M., NAKAMURA, S., SHIKANO, K., et al. Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan (E)*, 1990, vol. 11, no. 2, p. 71–76. DOI: 10.1250/ast.11.71

[11] TODA, A., BLACK, W., TOKUDA, K. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, vol. 15, no. 8, p. 2222–2235. DOI: 10.1109/TASL.2007.907344

[12] DESAI S., RAGHAVENDRA, E. V., YEGNANARAYANA, B., et al. Voice conversion using artificial neural networks. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Taipei (Taiwan), 2009, p. 3893–3896. DOI: 10.1109/ICASSP.2009.4960478

[13] SUNDERMANN, D., NEY, H. VTLN-based voice conversion. In *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. Darmstadt (Germany), 2003, p. 556–559. DOI: 10.1109/ISSPIT.2003.1341181

[14] ERRO, D., MORENO, A., BONAFONTE, A. Voice conversion based on weighted frequency warping. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, vol. 18, no. 5, p. 922–931. DOI: 10.1109/TASL.2009.2038663

[15] SERIZEL, R., GIULIANI, D. Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. In *IEEE Spoken Language Technology Workshop (SLT)*. South Lake Tahoe (NV, USA), 2014, p. 135–140. DOI: 10.1109/SLT.2014.7078563

[16] ERRO, D., NAVAS, E., HERNAEZ, I. Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, vol. 21, no. 3, p. 556–566. DOI: 10.1109/TASL.2012.2227735

[17] RABINER, L. R., SCHAFER, R. W. *Theory and Applications of Digital Speech Processing*. London (UK): Prentice Hall, 2011. ISBN: 9780136034285

[18] SIGMUND, M. Speaker discrimination using long-term spectrum of speech. *Information Technology and Control*, 2019, vol. 48, no. 3, p. 446–453. DOI: 10.5755/j01.itc.48.3.21248

[19] SIGMUND, M. Spectral analysis of speech under stress. *International Journal of Computer Science and Network Security*, 2007, vol. 7, no. 4, p. 170–172. ISSN: 1738-7906

[20] RUJZL, M., SIGMUND, M. Speech depersonalization based on the long-term spectrum of voice. *Authorea*, 2022, (preprint). DOI: 10.22541/au.166436464.40383121/v1

[21] TAGHAVI, S. M., MOHAMMADKHANI, G., JALILVAND, H. Speech intelligibility index: A literature review. *Auditory and Vestibular Research*, 2022, vol. 31, no. 3, p. 148–157. DOI: 10.18502/avr.v31i3.9861

[22] ANSI. *Methods for Calculation of the Speech Intelligibility Index*. ANSI S3.5-1997 [R2007]. New York, 2007.

[23] DRUGMAN, T., ALWAN, A. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Florence (Italy), 2011, p. 1973–1976. DOI: 10.21437/Interspeech.2011-519

[24] NOURTEL, H., CHAMPION, P., JOUVET, D., et al. Evaluation of speaker anonymization on emotional speech. In *ISCA Symposium on Security and Privacy in Speech Communication (SPSC)*. 2021, p. 62–66. DOI: 10.21437/SPSC.2021-13

[25] SIGMUND, M., DOSTAL, T. Analysis of emotional stress in speech. In *Proceedings of International Conference on Artificial Intelligence and Applications*. Innsbruck (Austria), 2004, p. 317–322.

[26] AALTO UNIVERSITY. *AALTO APARAT*. [Online] Cited 2023-06-15. Available at: research.spa.aalto.fi/projects/aparat/

[27] STANEK, M., SIGMUND, M. Psychological stress detection in speech using return-to-opening phase ratios in glottis. *Elektronika ir Elektrotechnika*, 2015, vol. 21, no. 5, p. 59–63. DOI: 10.5755/j01.eee.21.5.13336

[28] PRIBIL, J., PRIBILOVA, A., MATOUSEK, J. Evaluation of speaker de-identification based on voice gender and age conversion. *Journal of Electrical Engineering*, 2018, vol. 69, no. 2, p. 138–147. DOI: 10.2478/jee-2018-0017

About the Authors ...

Miroslav RUJZL received his M.Sc. at Brno University of Technology in 2021. His research interests include study and modeling of nonlinear dynamic systems with chaotic solution. In addition, he focuses on signal processing and embedded systems based on LPWAN networks.

Milan SIGMUND received his M.Sc. degree in Biomedical Engineering and the Ph.D. degree in Speech Signal Processing, both from Brno University of Technology. Currently, he is a Professor at the Dept. of Radio Electronics, Faculty of Electrical Engineering and Communication, Brno University of Technology.