Computer Speech & Language 65 (2021) 101118

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Analysis of gender and identity issues in depression detection on de-identified speech

Paula Lopez-Otero*[a], Laura Docio-Fernandez[b]

[a] Universidade da Coruña - CITIC, Facultad de Informática Campus de Elviña S/N, A Coruña, 15071 Spain
[b] AtlanTTic Research Center, E.E. Telecomunicación Campus Universitario S/N, Vigo, 36310 Spain

ARTICLE INFO

Article History:
Received 2 July 2019
Revised 13 April 2020
Accepted 18 May 2020
Available online 7 June 2020

Keywords:
Speaker de-identification
Depression detection
Voice conversion
Frequency warping and amplitude scaling
Generative adversarial networks

ABSTRACT

Research in the area of automatic monitoring of emotional state from speech permits envisaging future novel applications for the remote monitoring of some common mental disorders, such as depression. However, these tools raise some privacy concerns since speech is sent via telephone or the Internet, and it is further stored or processed in remote servers. Speaker de-identification can be used to protect the privacy of these patients, but this procedure might affect the ability to perceive the disease when using automatic depression detection approaches. It is also important that the resulting de-identified speech has enough quality since practitioners may need to listen to the recordings to assess the patients' state. This paper performs an extensive analysis of depression detection from de-identified speech using different de-identification approaches based on voice conversion. In previous work, a de-identification technique based on pretrained transformation functions was assessed in the context of depression detection. That strategy is speaker-independent (i.e. not speaker-specific) and gender-independent (i.e. the gender of the speaker is not necessarily preserved), which makes it possible to implement it in a real-world scenario where no parallel training data is required between input and source speakers. This paper aims at analyzing different aspects of the aforementioned speaker de-identification approach in a depression detection scenario: 1) compare the performance of the proposed speaker-independent technique with a speaker-dependent setting where parallel data between input and source speaker are available; 2) analyze how this system behaves when the gender of the speaker is preserved, since this might be a desirable feature and has not been addressed in previous work; 3) assess the performance of two different voice conversion methods in a setting where a limited amount of training data is available; specifically de-identification based on frequency warping and amplitude scaling (FW+AS) was compared with a strategy based on generative adversarial networks (GAN). Experimental validation was carried out in the framework of the Audio/Visual Emotion Challenge 2014, and the results suggest that speaker-independent and gender-dependent de-identification is the most suitable option for depression level estimation since the trade-off between de-identification and depression estimation performances was superior to the other alternatives. In addition, the results suggest that the de-identification approach based on GAN achieves better de-identification performance than FW+AS while achieving comparable results for depression detection.

*Corresponding author.
E-mail address: paula.lopez.otero@udc.gal (P. Lopez-Otero).

https://doi.org/10.1016/j.csl.2020.101118
0885-2308/

1. Introduction

Depression is a common mental disorder that causes people to experience depressed mood, loss of interest or pleasure, decreased energy, feelings of guilt or low self-worth, disturbed sleep or appetite, and poor concentration (Marcus et al., 2012). Current tendencies in mood disorders tend to avoid inpatient treatments (i.e. long-term stays at hospitals or institutions) in favor of outpatient treatments, since the latter promotes the inclusion of the patients in the society. These treatments are highly demanding for practitioners, as they require constant supervision in order to detect changes in the patients' state that might lead to suicidal behaviors (Hor and Taylor, 2010; Cummins et al., 2015a). This, along with the prevalence of depression in nowadays society, creates the need for tools that allow the monitoring of the mental health state of patients suffering this disorder. The current available technology makes possible to automatically analyze the emotional state (Karam et al., 2014; Gravenhorst et al., 2015; Ben-Zeev et al., 2015). This is done by means of different biomarkers that carry information about depression, such as the voice (Sturim et al., 2011), facial expression (Cohn et al., 2009), eye gaze, head pose (Alghowinem et al., 2017) or features extracted from brain magnetic resonance images (MRI) (Kipli and Kouzani, 2013; Kipli et al., 2013). Among these biomarkers, speech is the most useful for developing this kind of tools, since it can be measured cheaply, remotely, non-invasively and non-intrusively Cummins et al. (2015a).

Remote monitoring of the emotional state using speech has some drawbacks, among which the privacy issue can be highlighted. Sending speech through the Internet has a potential risk of having this information intercepted by attackers. Also, the storage or processing of these recordings on proprietary or third-party cloud-based infrastructures is common, which enables speech data to be exposed and intercepted by fraudsters or hackers raising serious privacy issues. The interception of such data constitutes an undeniable intrusion to personal privacy, as the information used in speaker characterization can also be misused by fraudsters for other purposes. For example, a user's speech data may be acquired and used to attack speaker verification systems providing access to personal or sensitive data. In addition, speech also has personal and private information about emotional and health status that most of us would not voluntarily trust others. These risks can discourage the use of apps or web interfaces for monitoring mental disorders to avoid the exposure of sensitive information to untrusted individuals. A possible solution for this privacy concern consists in applying de-identification, which is a process by which a data custodian alters or removes identifying information from a dataset, making it harder for users of the data to determine the identity of the data subjects (Garfinkel, 2015). In the case of systems dealing with speech, speaker de-identification implies modifying the voice characteristics in order to achieve a new spoken document in which it is not possible to recognize the identity of the speaker. The most common speaker de-identification techniques are either not reversible or robust to attacks when the conversion parameters are not available, which makes them suitable for privacy preservation in speech-related applications. Besides this, the quality of the de-identified recordings is of paramount importance since practitioners may be interested in listening to the recordings to assess the patients' mental state; in addition, removing the identifying information from these spoken data would also ease the process of compiling speech for research purposes in this field since the patients would not feel so exposed when collaborating in this type of studies.

Speaker de-identification based on voice conversion modifies the voice characteristics, which might also result in partial removal or alteration of the depression related information. The influence of speaker de-identification in depression detection has been assessed in previous work (Lopez-Otero et al., 2017b), and the results showed that speaker de-identification did not affect depression detection techniques under some experimental conditions. Nevertheless, the focus has been placed on speaker-independent de-identification techniques, i.e. approaches where voice conversion functions can be applied regardless the identity of the input speaker (Magariños et al., 2016; 2017). The use of such strategies was motivated by the need for a parallel corpus between the source and target speaker or, equivalently, the lack of a speech corpus for depression detection that included parallel corpora that allowed the training of voice conversion functions. In addition, the proposed approaches implied either a mandatory (Magariños et al., 2016) or implicit (Magariños et al., 2017) change of speaker gender since, in the latter approach, the speaker was converted to the most dissimilar voice, which usually belongs to the opposite gender. This change of gender makes it difficult to use gender-dependent depression detection approaches since there is uncertainty about the actual gender of the speaker. It is well-known that gender plays an important role in psychological illnesses such as depression (Nolen-Hoeksema, 1987); also, gender-dependent depression detection systems usually exhibit a better performance according to the literature (Huang et al., 2016; Pampouchidou et al., 2016).

This paper aims at analyzing the impact of different speaker de-identification techniques on automatic depression detection. These techniques are based on pre-trained voice conversion functions (Magariños et al., 2017) since this strategy allows the de-identification of any speaker regardless the existence of a parallel corpus. Specifically, the focus of this paper is placed on three aspects: (1) training speaker-dependent voice conversion functions versus using speaker-independent approaches; (2) performing gender-dependent de-identification (i.e. the source gender is equal to the target gender) versus gender-independent de-identification; (3) comparing speaker de-identification techniques based on frequency warping plus amplitude scaling (FW+AS), as in Magariños et al. (2017), and on generative adversarial networks (GAN) (Saito et al., 2018b). This analysis was performed in the framework of the AVEC 2014 depression sub-challenge (Valstar et al., 2014).

The rest of this paper is organized as follows: Section 2 reviews the most common speaker de-identification and depression detection techniques in the literature; Section 3 describes the speaker de-identification techniques applied in this paper; Section 4 presents the depression detection approach used in this work; Section 5 overviews the experimental setup used to assess depression detection on de-identified speech, and presents the experimental results; Section 8 discusses the achieved results; and finally Section 9 summarizes the conclusions and future work.

2. Related work

This section reviews the most common techniques for speaker de-identification and depression detection, and explains the design decisions adopted in this paper.

2.1. Speaker de-identification

There are two main different strategies to perform speaker de-identification: one of them consists in performing automatic speech recognition (ASR) followed by a text-to-speech (TTS) system (Justin et al., 2015), while the other consists in using voice conversion techniques (Jin et al., 2009; Abou-Zleikha et al., 2015; Magariños et al., 2016; 2017). The use of voice conversion for speaker de-identification is more extended for several reasons: the ASR+TTS approach removes information that has been proven to be an indicator of depression, such as prosody, intonation or speech rate (Cummins, 2016; Williamson et al., 2014; Lopez-Otero et al., 2014a; 2014b); it requires the availability of ASR and TTS modules for a given language; and the quality of the resulting speech strongly depends on the performance of the transcription stage. Hence, the most common techniques for speaker de-identification rely on voice conversion, either applying frequency warping techniques (Sündermann and Ney, 2003; Erro et al., 2010; Zorila et al., 2012; Erro et al., 2013) or approaches based on deep learning (Mohammadi and Kain, 2017; Bahmaninezhad et al., 2018; Saito et al., 2018b). However, both strategies share the limitation of requiring a parallel corpus between the source and target speakers to train the conversion parameters, even though research on non-parallel methods for voice conversion has recently gained attention of the research community and some promising results have been presented (Hsu et al., 2016; Lorenzo-Trueba et al., 2018; Saito et al., 2018a; Kaneko and Kameoka, 2018; Zhang et al., 2020). However, there is still a gap between the quality of parallel and non-parallel VC (Lorenzo-Trueba et al., 2018), and non-parallel approaches usually need larger training corpora for training. This motivated the use of parallel methods in this paper and, since parallel data is difficult to collect in practice, we aim at facing the challenge of how to successfully perform voice conversion with limited parallel data.

In this paper, speaker de-identification based on pre-trained conversion functions Magariños et al. (2017) is chosen since it allows the training of speaker/gender-dependent/independent functions, which is the main focus of these experiments. In addition, besides the FW+AS voice conversion strategy described in Magariños et al. (2017), the GAN-based technique described in Saito et al. (2018b) is evaluated in this paper in order to assess the performance of depression detection on speech that is de-identified using a deep learning approach. This GAN-based strategy was chosen among others because it exhibited good de-identification performance in preliminary experiments.

2.2. Depression detection

Speech is a biomarker widely used in depression detection because it conveys information both in its lexical content (Williamson et al., 2016; Correia et al., 2016; Lopez-Otero et al., 2017a) and in its acoustic and prosodic characteristics (Williamson et al., 2014; Lopez-Otero et al., 2015; Morales and Levitan, 2016). In addition, research on this field has been largely boosted by the organization of the Audio/Visual Emotion challenges (AVEC) (Valstar et al., 2013; 2014; 2016; Ringeval et al., 2017), where a common evaluation framework is provided to the participants. Speech-based depression detection using either lexical content or acoustic and prosodic characteristics lead to acceptable results, but the advantage of those methods based on acoustic characteristics is that they do not require any knowledge about the language being spoken and, therefore, no specific linguistic resources are necessary for system training. Hence, in this paper, depression detection is carried out using acoustic and prosodic characteristics of speech.

Most of the depression detection strategies based on speech features found in the literature follow the conventional feature extraction stage followed by supervised model training and classification. In particular, the use of GMMs for depression detection is still quite common (Cummins et al., 2015b; Syed et al., 2017), but i-vectors have become more usual in recent works (Mitra et al., 2014; Lopez-Otero et al., 2015; Nasir et al., 2016). Another approach that was reported to exhibit good performance consists in using the correlation structure of different features followed by a Gaussian staircase model for depression level estimation (Williamson et al., 2014; 2016). The use of a large set of functionals extracted from low level descriptors is quite extended, either followed by an SVM (Nasir et al., 2016) or deep learning approaches (Yang et al., 2017) for classification. Formants, prosodic and voice quality features are also quite widespread (Yang et al., 2016; Sun et al., 2017), given that these representations are usually encouraged by their use as part of the AVEC evaluation baselines (Valstar et al., 2013; 2014; 2016; Ringeval et al., 2017). Also, in the late years, the popularization of deep learning has led to depression detection approaches using long-short term memory (LSTM) networks with high level features as input (Alhanai et al., 2018) or convolutional neural networks (CNNs) with raw speech and spectrograms as input (He and Cao, 2018), to cite some examples. A recent paper (Correia et al., 2018) compares classifiers based on SVMs with eGeMAPS features (Eyben et al., 2016) and CNNs using spectrograms for depression detection, and the results showed a superior performance of the SVM classifier.

In this work, a depression detection approach based on i-vectors and SVM for regression was adopted from previous work Lopez-Otero et al. (2015, 2017b) since it exhibited a good performance for depression detection in previous research and, in addition, the literature does not suggest a superior performance of deep learning-based approaches.

3. Speaker de-identification

As mentioned in Section 2, speaker de-identification is usually carried out by means of voice conversion techniques, which modify the voice characteristics of a source speaker in order to make it sound like a different target speaker. This modification is achieved by applying a transformation that converts the source speaker into the target speaker, and its parameters must be trained using speech of the source and target speakers. Since these speech utterances are not always available (especially when a parallel corpus is required for training), the speaker de-identification technique based on pre-trained transformation functions proposed in Magariños et al. (2017) is used in this paper. This section first describes this technique and how to apply it in speaker/gender-dependent/independent settings, and then the different voice conversion approaches used for de-identification in this work are described, namely frequency warping plus amplitude scaling and generative adversarial networks. These two methods were chosen in order to assess the validity of two different voice conversion paradigms (parametric and generative, respectively) for depression detection on de-identified speech: parametric methods have already been evaluated in the literature (Lopez-Otero et al., 2017b), but the use of generative methods for this task is a novelty.

3.1. Speaker de-identification based on pre-trained voice conversion functions

A typical voice conversion system considers two different speakers, namely source and target, and the aim of voice conversion is making the source speaker sound like the target speaker. For this purpose, a voice conversion function from *source* to *target* is trained. However, when using voice conversion techniques for speaker de-identification in a real-world setting, it is unlikely that there are data from the speaker to be de-identified available for training the conversion functions. This problem can be overcome using the de-identification strategy based on pre-trained conversion functions (Magariños et al., 2017). In this paradigm, three different speakers are considered: input (i.e. the speaker to be de-identified), source and target. The idea is simple: voice conversion is applied to the input speaker using a conversion function trained using the source and target speakers. To apply this technique, the most suitable source and target speakers must be chosen depending on the input speaker to be de-identified.

The motivation behind this method lies in the premise that the objective of de-identification is not mimicking a target speaker but producing speech in which the identity of the speaker is not recognizable. Hence, choosing a source speaker that is similar to the input speaker ensures that the transformation parameters are as suitable as possible for the input speaker. In addition, selecting the most dissimilar target speaker maximizes the chance of achieving speech that sounds very different to the input speaker. The main advantage of this method is the possibility of de-identifying speech from any speaker regardless of the availability of a parallel corpus including data from this input speaker.

Formally, given an input speaker S_{input}, a set of n_s potential source speakers $S_{source} = \{S_{source_1}, \ldots, S_{source_{n_s}}\}$ and a set of n_t potential target speakers $S_{target} = \{S_{target_1}, \ldots, S_{target_{n_t}}\}$, a source speaker S^*_{source} is selected such that it is the most similar to S_{input}, and and a target speaker S^*_{target} is selected such that it is the most dissimilar to S_{source}, as described in Magariños et al. (2017).

In this work, the similarity between speakers is obtained using the i-vector paradigm (Dehak et al., 2010) combined with probabilistic linear discriminant analysis (PLDA) scoring (Garcia-Romero and Espy-Wilson, 2011). This technique defines a low dimensional space, namely total variability space, in which speech segments are represented by a vector of total factors, commonly known as i-vector (Dehak et al., 2010). Given a speech utterance, its corresponding GMM supervector $\mathbf{M}$ can be decomposed as:

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

where $\mathbf{m}$ is the speaker and channel-independent supervector, $\mathbf{T}$ is a low rank total variability matrix, and $\mathbf{w}$ is the i-vector corresponding to the GMM supervector. A pair of i-vectors can be compared by computing their PLDA scoring (Garcia-Romero and Espy-Wilson, 2011).

Given an input speaker S_{input}, its i-vector $\mathbf{w}_{input}$ is extracted and compared with all the potential source speakers, selecting the one that maximizes the PLDA score:

$$S^*_{source} = \arg\max_{i \in 1,\ldots,n_s} \text{score}(\mathbf{w}_{input}, \mathbf{w}_{source_i}) \tag{2}$$

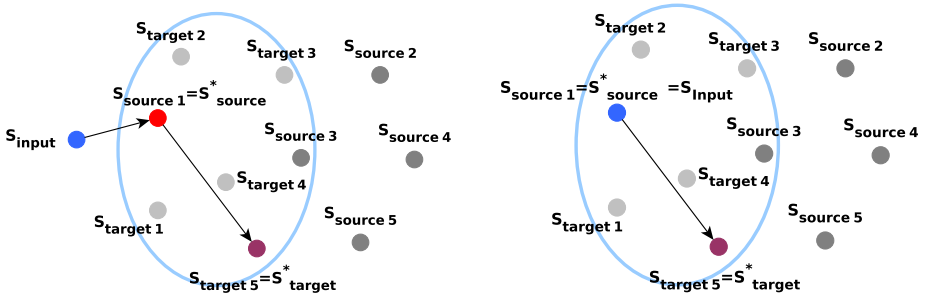
Equivalently, the target speaker is selected by minimizing the PLDA score between the source speaker and the potential target speakers:

$$S^*_{target} = \arg\min_{i \in 1,\ldots,n_t} \text{score}(\mathbf{w}^*_{source}, \mathbf{w}_{target_i}) \tag{3}$$

This method can be used to define speaker dependent/independent and gender dependent/independent de-identification strategies by setting restrictions to the sets of source and target speakers.

3.1.1. Speaker-independent versus speaker-dependent de-identification

The de-identification strategy proposed in Magariños et al. (2017) is speaker-independent, i.e. it is assumed that the input speaker does not belong to the set of source speakers, likely because there are not parallel corpora available from the input speaker. When parallel corpora are available, it is possible to transform this approach into a speaker-dependent method by considering that $S^*_{source} = S_{input}$ and then selecting the target speaker as described in Eq. (3). The speaker-independent and speaker-dependent de-identification methods are depicted in Figs. 1a and 1 b, respectively.



(a) Gender-independent and speaker-independent strategy (b) Gender-independent and speaker-dependent strategy

Fig. 1. Diagram of the proposed gender-independent selection strategies for voice conversion. The dots represent the different speakers, and those inside the blue ellipse are training speakers that share a parallel corpus. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.1.2. Gender-independent versus gender-dependent de-identification

When describing the speaker-independent/dependent de-identification strategies, no mention was made about the nature of the available source and target speakers. In fact, in Magariños et al. (2017), the potential source and target speakers could be either male or female speakers regardless the gender of the input speaker. Nevertheless, performing gender-dependent de-identification can be interesting: first, the gender of the original speaker is preserved, which allows the use of depression detection methods that take gender into consideration; second, the quality of the resulting speech is probably better since the source and target voices are more similar between them, which leads to a slighter modification of the cepstral parameters.

A gender-dependent version of the two aforementioned de-identification strategies can be easily achieved: first the gender $g \in \{\text{male, female}\}$ of the input utterance is detected, and then the source and target speakers are obtained from the sets S^g_{source} and S^g_{target} that include source and target speakers, respectively, of gender g. Figs. 2a and 2 b depict the speaker-independent and speaker-dependent versions, respectively, of this gender-dependent strategy.

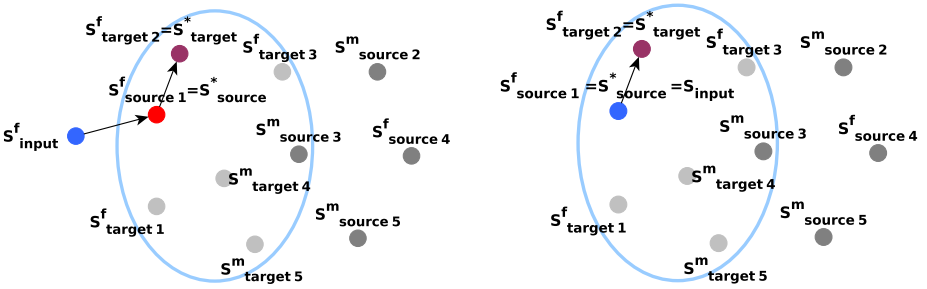
3.2. Voice transformation based on frequency warping and amplitude scaling

One of the methods used in this work for speaker de-identification is based on frequency warping with amplitude scaling (FW +AS) voice conversion. Regardless the increasing popularity of DNN approaches for voice conversion, these methods are still used in the literature because they can achieve good results with little training data (Gao et al., 2019; Zhao et al., 2019b).

FW+AS can be implemented as an affine transformation in the cepstral domain (Pitz and Ney, 2005; Erro et al., 2015):

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{4}$$

where **x** and **y** are the original and transformed Mel-cepstral (MCEP) vectors, respectively; and **A** and **b** are the FW and AS parameters, respectively. Matrix A can be trained by means of a dynamic frequency warping (DFW) technique (H. Valbret et al., 1992) so



(a) Gender-dependent and speaker-independent strategy (b) Gender-dependent and speaker-dependent strategy

Fig. 2. Diagram of the proposed gender-dependent speaker selection strategies for voice conversion. The dots represent the different speakers, and those inside the blue ellipse are training speakers that share a parallel corpus. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

that the source spectra and the target spectra are maximally close. In this work the training procedure described in (Magariños et al., 2017) was used to obtain the FW parameters. First, matrix **A** is computed as:

$$\mathbf{A} = \mathbf{C} \cdot \mathbf{W} \cdot \mathbf{S} \tag{5}$$

where **S** is a rectangular matrix that transforms a Mel-cepstral vector into its corresponding discrete log-amplitude spectrum, **C** is its pseudo-inverse, and **W** is a sparse square matrix that captures the correspondence between original and warped spectral bins, and is obtained through the DFW technique. Then, the AS vector is computed as the cepstral difference between the averaged target vectors $\bar{\mathbf{y}}$ and the warped version of the averaged source vectors $\bar{\mathbf{x}}$:

$$\mathbf{b} = \bar{\mathbf{y}} - \mathbf{A}\bar{\mathbf{x}} \tag{6}$$

As shown in Magariños et al. (2016), performing adaptation of the fundamental frequency (F0) on top of FW+AS transformation dramatically improves de-identification performance. Hence, in this work, a mean and variance adaptation of the F0 is applied:

$$\log \hat{f}_0^t = \frac{\sigma_t}{\sigma_s}(\log f_0^s - \mu_s) + \mu_t \tag{7}$$

where f_0^s represents the frame-level values of the *source* speaker's F0; μ_s, σ_s, μ_t, σ_t are the mean and standard deviation of the *source* and *target* F0 in the log domain, respectively; and $\hat{f}_0^t$ is the adapted F0 of the *target* speaker.

In this system, Ahocoder (Erro et al., 2011) was used to extract 40 MCEP coefficients, F0 and band aperiodicity (BAP) features. Given a utterance to de-identify, its MCEP coefficients are converted applying the FW and AS parameters described above, and then the F0 is scaled following Eq. (7). Finally, Ahocoder is used to synthesize the de-identified speech utterance using these transformed features and the original BAP features, which remain unchanged.

3.3. Generative adversarial network

Generative adversarial networks (GAN) are a very popular approach for speech synthesis and voice conversion, either in parallel (Sisman et al., 2018) and non-parallel (Paul et al., 2019; Chen et al., 2019) settings. The success of this technique is leading to the investigation of different variants to further improve their performance (Zhao et al., 2019a; Chen et al., 2019). The system proposed in Saito et al. (2018b) for both speech synthesis and voice conversion, and is used in this work as a deep learning approach for speaker de-identification. In this strategy, GANs are used to avoid the oversmoothing effect observed in other DNN-based strategies. The GAN combines two competing neural networks: a discriminator and a generator. The discriminator tries to distinguish natural and generated samples, while the generator aims at deceiving the discriminator.

Formally, given a set of T samples $\mathbf{x}_t$ of the training speaker, they are transformed into samples $\mathbf{y}_t$ that sound as spoken by the target speaker. This is done by iteratively training the generator and discriminator DNNs using a stochastic gradient descent optimizer. First, the loss of the discriminator is computed as

$$L_D^{(GAN)}(\mathbf{x}, \mathbf{y}) = -\frac{1}{T}\sum_{t=1}^{T} \log \frac{1}{1 + \exp\left(-D(\mathbf{x}_t)\right)} - \frac{1}{T}\sum_{t=1}^{T} \log\left(1 - \frac{1}{1 + \exp\left(-D(\mathbf{y}_t)\right)}\right) \tag{8}$$

Then, the parameters of the discriminator are updated and then the adversarial loss of the generator is computed as

$$L_{ADV}^{(GAN)}(\mathbf{y}) = -\frac{1}{T}\sum_{t=1}^{T} \log \frac{1}{1 + \exp\left(-D(\mathbf{y}_t)\right)} \tag{9}$$

The parameters of the generator are subsequently updated. The loss function used for training the GAN is

$$L_G(\mathbf{x}, \mathbf{y}) = L_{MGE}(\mathbf{x}, \mathbf{y}) + \omega_D \frac{E_{L_{MGE}}}{E_{L_{ADV}}} L_{ADV}^{(GAN)}(\mathbf{y}) \tag{10}$$

where

$$L_{MGE}(\mathbf{x}, \mathbf{y}) = \frac{1}{T}(\mathbf{y} - \mathbf{x})^T(\mathbf{y} - \mathbf{x}) \tag{11}$$

is the loss of the minimum generator error trainer as described in Saito et al. (2018b) and ω_D is a hyper-parameter that controls the weight given to the loss function of the GAN network.

In this system, speech analysis and synthesis were performed using Ahocoder instead of WORLD Morise et al. (2016) and SPTK[1] as described in Saito et al. (2018b). This design decision was made to enable a fair comparison between FW+AS and GAN methods; in addition, informal listening tests did not show a significant difference between de-identified speech using one set of features or the other. As in Section 3.2, the training procedure described above is used to transform the MCEP features, while the F0 is scaled as described in Section 3.2 and BAP features remain unchanged. The architecture of the network is as follows: the

[1] http://sp-tk-sourceforge.net

generator and discriminator have hidden layers with 3x512 units and 2x256 units, respectively, and the learning rate is 0.01 for both of them. In these experiments, ω_D was set to 1.

4. Depression detection

The depression detection strategy employed in this work is based on the representation of speech information using i-vectors (Cummins, 2016; Lopez-Otero et al., 2014a; 2015; 2017b). The i-vector paradigm allows for a compact low-dimensional representation of the different sources of variability present in a speech recording, such as gender, age, channel, speaker or message, that can be better compensated or modeled in a latter stage. An overview of the system is presented in Fig. 3, where three blocks are clearly distinguishable: front-end, speech representation and depression level estimation.

First, features extracted from the waveform should preserve the information about the depressive state of the patient. Different features have been proposed in the literature for depression detection, such as those described in Cummins (2016), Williamson et al. (2014), Lopez-Otero et al. (2014a), Lopez-Otero et al. (2014b), but there is no agreement about which are the most suitable for this task. Following Lopez-Otero et al. (2015, 2017b), 13 perceptual linear prediction (PLP) cepstral coefficients combined with two pitch-related features (F0 and voicing probability) where used in this work. These features achieve a speech representation that conveys both spectral and prosodic information such as rhythm or intonation, which is embedded in the fundamental frequency. The features were extracted every 10 ms using a 25 ms window, and they were augmented with their delta and acceleration coefficients in order to capture short-term dynamics, which led to feature vectors of dimension 45. Cepstral mean subtraction was performed to reduce the influence of recording channel and noise. Voice activity detection (VAD) was applied before feature extraction in order to exclude silence regions since their spectral content do not carry information about the depressive state of the speaker. Specifically, the VAD algorithm described in Basu (2003) was used.

Once features are extracted from the training recordings, they are used to train a Universal background model (UBM) following the expectation-maximization (EM) algorithm (Reynolds et al., 2000) since it is necessary for extracting i-vectors as defined in Eq. (1). Then, given the UBM and a set of training speech utterances, the total variability matrix **T** can be trained as described in Dehak et al. (2010). The i-vectors extracted using the trained **T** matrix are also length-normalized as described in Garcia-Romero and Espy-Wilson (2011) to Gaussianize their distribution.

In order to obtain several speaker turns, both for training and testing, sessions are usually split following some criteria such as windowing, dividing the recordings by sentences or using the VAD output to split the recording in points where the speaker makes pauses. In this work, the recordings were segmented using a sliding window of 20 s with 15 s overlap as in Lopez-Otero et al. (2017b).

Finally, given the i-vectors of the speaker turns and the depression level corresponding to each i-vector, a regressor to estimate the depression level of each utterance can be trained. In this work, a support vector regressor (SVR) with linear kernel was used for that purpose as in Lopez-Otero et al. (2015, 2017b). Since each session is usually formed by several speech segments, the overall decision for a whole session is taken by computing the average value of the depression levels predicted for each speech segment.

5. Experimental setup

Several speech datasets and evaluation protocols were used to assess the effectiveness of the depression level estimation system and the speaker de-identification approaches described in Section 3. This section describes the principal characteristics of such datasets and how them are used.

5.1. Speech corpora

Table 1 summarizes the use of each speech corpus, which are described in detail below.

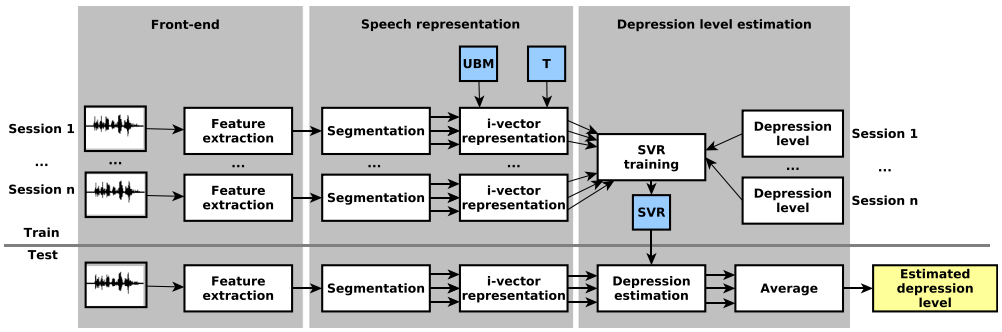


Fig. 3. Block diagram of the depression level estimation system.

Table 1
Databases used for each task performed in the experiments.

Task	Train	Test
Depression detection	Biosecure (UBM, total variability matrix)	AVEC 2014
Speaker-independent de-identification	VCTK	AVEC 2014
Speaker-dependent de-identification	VCTK AVEC 2013	AVEC 2014
Evaluation (speaker verification)	Biosecure (UBM, total variability matrix, PLDA transform)	AVEC 2014

5.1.1. AVEC 2014

The recordings of AVEC 2014 depression recognition sub-challenge (Valstar et al., 2014) were used in this paper to evaluate the performance of both the speaker de-identification system and the depression level estimation system. These recordings are a subset of the audio-visual depressive language corpus (AVDLC) (Valstar et al., 2013), which consist in a series of PowerPoint-guided tasks: reading of excerpts of novels and fables, singing, reminiscing, making up a story applying the Thematic Apperception Test (TAT) and sustained vowel phonation. In the AVEC 2014 data, only two of the tasks were included: reading an excerpt of the fable "Die Sonne und der Wind" and free responses to questions such as "What is your favourite dish?" or "Discuss a sad childhood memory" in German, which is the native language of the speakers.

The sessions included in these databases were recorded in quiet locations using a laptop and a headset. Each recording features either a male or female speaker, who may have made several recordings with a time interval of two weeks between them. The subjects' ages ranged between 18 and 63 years (mean 31.5). Not all of the recorded sessions in AVEC 2014 were used in these experiments, only those that allowed the training of speaker-dependent conversion functions were selected (as further explained in the next subsection). This led to a set of 120 recordings uttered by 65 speakers: 25 speakers recorded only one session, 25 recorded two sessions, and 15 recorded three sessions.

The depression level corresponding to each session was provided with the audio recordings. This level was obtained using the Beck's depression inventory (BDI-II) (Beck et al., 1996), which is a self-assessment questionnaire in which the patient must answer questions related to depression symptoms such as irritability, hopelessness, fatigue or lack of interest in sex. It consists of 21 questions whose answers are rated in a scale from 0 to 3 and, when summed, they yield a depression level ranging from 0 to 63. During the capture of the database, the speakers were asked to fill in the BDI-II questionnaire after performing the aforementioned tasks.

5.1.2. VCTK

The VCTK Corpus (Veaux et al., 2016) was used to train de-identification functions. Specifically, the speakers included in this dataset were used as source and target speakers for the speaker-independent de-identification approach, and as target speakers for the speaker-dependent strategy. This corpus comprises speech uttered by 109 native English speakers. Each speaker was asked to read about 400 sentences which were not the same for all of them. Nevertheless, all the speakers read "The Rainbow Passage" (Fairbanks, 1960), so those 19 sentences were used as a parallel corpus to train the conversion functions. This corpus amounts to around 2 min of speech per speaker, which is a reasonable amount of speech for training conversion functions as seen in previous work (Magariños et al., 2017; Lopez-Otero et al., 2017b).

5.1.3. AVEC 2013

Some speakers included in the recordings of the AVEC 2013 depression recognition sub-challenge Valstar et al. (2013) were used as source speakers for the speaker-dependent de-identification strategy. One of the tasks included in the protocol defined for the AVDLC corpus included reading "The Rainbow Passage", which makes it possible to establish a parallel corpus between these speakers and those in VCTK Corpus. Hence, those speakers that appeared both in AVEC 2013 and 2014 data and successfully completed this task were included in the experiments, leading to the aforementioned set of 65 speakers. To be able to perform this training, the recordings of "The Rainbow Passage" in AVEC 2013 were manually split into the same 19 utterances as in the VCTK corpus. It must be mentioned that, in general, voice conversion functions are trained using data that was specifically recorded for this purpose, so the speaker-dependent de-identification strategy assessed in this work is more challenging because the recordings are spoken by non-native speakers of English in recording conditions that were not ideal.

5.1.4. Biosecure

In addition to the aforementioned datasets, Biosecure database Ortega-Garcia et al. (2010) was also used in the experiments. Specifically, the DS1 partition of this multimodal database was used in the speaker verification experiments as well as in the depression level estimation module. It consists of around 18 h of audio of 316 different speakers from 7 different countries, acquired over the Internet under unsupervised conditions. This dataset was used to train UBMs, PLDA transformations and total variability matrices for all the tasks: selection of similar/dissimilar speakers in the speaker de-identification strategy; extraction

of i-vectors for depression level estimation, and speaker verification experiments. The number of Gaussians of the UBM and the dimension of the i-vectors was individually tuned for each specific task.

It must be mentioned that the use of this dataset does not have any impact on the experiments or the results: an i-vector extractor was necessary to perform the experiments and, instead of using out-of-the-box models, they were trained using these data.

6. Speaker de-identification experiments

This section presents the performance of the different de-identification strategies described in Section 3. To ease the understanding of this section, abbreviations for the de-identification techniques are defined in Table 2.

6.1. Evaluation metrics

To assess the quality of the proposed speaker de-identification approaches objective and subjective measures were used. First, speaker de-identification was assessed by means of speaker verification experiments as in Lopez-Otero et al. (2017b). Specifically, a state-of-art speaker verification system was developed using the Kaldi toolkit (Povey et al., 2011), which uses i-vectors for speech representation combined with PLDA scoring (Garcia-Romero and Espy-Wilson, 2011). First, a speaker verification experiment with the original enrollment and test speech utterances (i.e. voice without applying de-identification) was performed to establish a reference for comparison with the results achieved on de-identified speech. Then, a speaker verification experiment using the original enrollment utterances and the de-identified test utterances was done. It is expected that the experiment on the original test utterances leads to a good speaker verification performance and the ones on de-identified utterances achieve a much worse performance since this would indicate that it is not possible to identify the speaker after de-identification. The performance of the aforementioned experiments was measured by means of the equal error rate (EER), accompanied by detection error trade-off (DET) curves. The EER represents the error of the working point at which the false rejection probability is equal to the false acceptance probability, while the DET curve is a plot of the false acceptance and false rejection probabilities at different decision thresholds.

In addition, as in Magariños et al. (2017), the degree of de-identification was also measured using the Mel-cepstral distortion (MCD): when comparing the MCEP features of the original and de-identified utterances, this measure reflects how far the de-identified vectors are from the original vectors:

$$MCD[dB] = \frac{1}{M}\sum_{m=1}^{M}\sqrt{2\sum_{d=1}^{24}(c_{m,d}-\widehat{c}_{m,d})^2} \tag{12}$$

where $c_{m,d}$ and $\widehat{c}_{m,d}$ are the dth MCEP feature of vector m of the original and de-identified utterances, respectively.

Unfortunately, the end-user licence agreement (EULA) of AVEC database explicitly prohibits sharing clips of the audio to non-signers of that document, which frustrated the possibility of performing a listening test with a reasonable number of participants. Nevertheless, informal listening tests were done by the authors of the paper. Specifically, the similarity test described in Wester et al. (2016) for the Voice Conversion Challenge 2016 was done: it consists in, given two samples of speech (original and de-identified), deciding whether the samples are from same or different speakers with two confidence levels (not sure, absolutely sure). In these experiments, the percentage of pairs labeled as "Different, absolutely sure" is expected to be high since this would mean that the speaker is correctly de-identified. The validity of this experiment is limited because of the small number of participants on the test, but previous research showed that the objective and subjective metrics used for assess speaker de-identification are correlated (Magariños et al., 2017).

6.2. Results

Fig. 4 depicts the DET curves obtained when performing the speaker verification experiment with the original and de-identified recordings.

Experiments were done on the AVEC 2014 data: first all the sessions of the same speaker were joined into a single one, and then this recording was divided into utterances of 15 s each. The first utterance of each speaker was used for enrollment while the remaining ones were used for testing. This led to a total of 620 client trials and 39,680 impostor trials.

Table 2
Abbreviations assigned to the different speaker de-identification strategies.

Name	Gender-dependent?	Speaker-dependent?
GI-SI	no	no
GI-SD	no	yes
GD-SI	yes	no
GD-SD	yes	yes

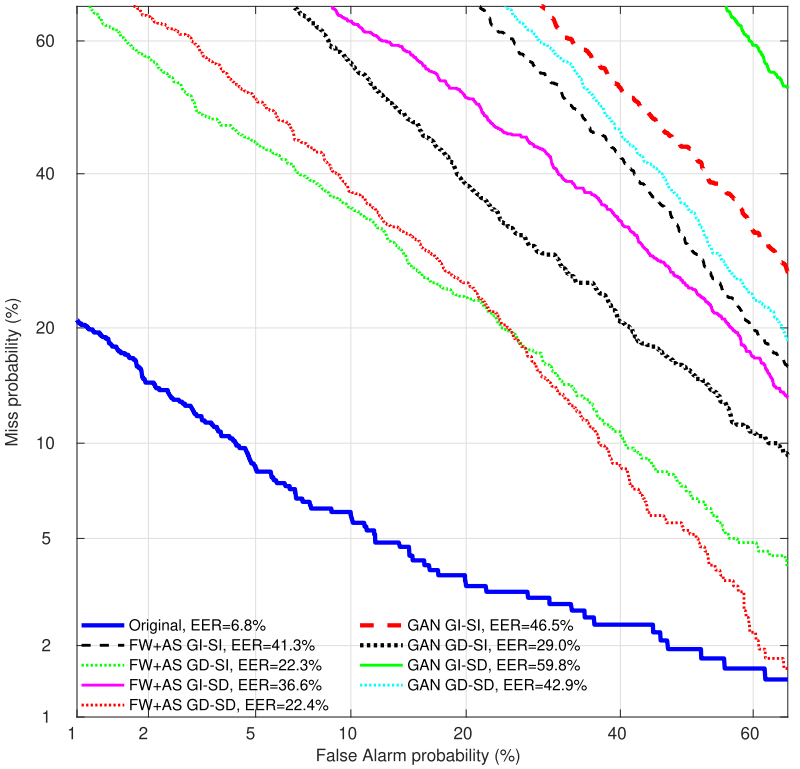


Fig. 4. DET curves of original and de-identified speech using different de-identification strategies based on frequency warping plus amplitude scaling (FW+AS) and generative adversarial networks (GAN). The de-identification strategies are: gender-independent and speaker-independent (GI-SI), gender-dependent and speaker-independent (GD-SI), gender-independent and speaker-dependent (GI-SD), gender-dependent and speaker-dependent (GD-SD). Results using the original recordings (i.e. without applying de-identification) are shown for comparison.

As shown, an EER of 6.8% is achieved when using the original utterances which, according to Sadjadi et al. (2017), is an acceptable result when dealing with unconstrained speech. With respect to the de-identified speech, the figure shows that the degree of de-identification achieved with the gender-dependent strategies is lower than that of gender-independent approaches except for the GAN-based GD-SD strategy, which achieves de-identification results comparable to those of gender-independent strategies. Nevertheless, the lowest EER was 22.3% (around four times the EER achieved with original speech), which is an acceptable de-identification result according to Lopez-Otero et al. (2017b). When using transforms based on FW+AS, the gender-dependent results where very similar for both GD-SD and GD-SI strategies, while GD-SD achieved an EER much higher than that of GD-SI for GAN-based transforms. In the case of gender-independent de-identification, the EER obtained with the GI-SD method was slightly lower than that of the GI-SI approach with the FW+AS transform, which means that the speaker-independent conversion leads to voices where recognizing the identity of the speaker is more difficult. The contrary occurred with the GAN-based de-identification strategy, where the speaker-dependent transform achieved a higher EER. In general, according to the results displayed in Fig. 4, speaker de-identification strategies based on GAN conversions achieve higher de-identification rates than those based on FW+AS in both gender-dependent and gender-independent settings.

It must be highlighted that the speaker-independent conversion functions were trained using source data that were not collected for voice conversion purposes. Therefore, some of the recordings were noisy, the speech was not as clear as desired, and some speakers did not have good pronunciation skills in English language. Nevertheless, an informal listening test showed that the resulting de-identified speech had the same quality as with the speaker-dependent conversion functions, which were trained using a corpus that is more suitable for this task.

As stated above, the different de-identification techniques were also evaluated in terms of their MCD when comparing the MCEP features of the original and de-identified utterances, and the results are shown in Table 3. In these experiments, the MCEP vectors of all the de-identified utterances (i.e. 120 AVEC sessions) were considered for computing the MCD.

According to this measure, gender-dependent de-identification based on FW+AS leads to MCEP vectors that are closer to the original vectors compared to gender-independent de-identification. In the case of GAN-based de-identification this difference is negligible. As expected from the DET analysis presented above, gender-dependent de-identification based on GAN achieves a higher MCD than FW+AS.

It would also be interesting to observe the MCD of the de-identified utterances compared to the target speaker, but this is not possible since MCD computation requires parallel utterances. Nevertheless, according to the equivalent values presented in Magariños et al. (2017), the MCD between target and de-identified utterances in a speaker-independent setting is very similar to that

Table 3
Average MCD (dB) with 95% confidence intervals between the original and de-identified utterances of AVEC 2014.

	MCD	
System	FW+AS	GAN
GI-SI	7.84 ± 0.28	6.57 ± 0.24
GI-SD	8.38 ± 0.37	7.85 ± 0.16
GD-SI	5.89 ± 0.26	6.12 ± 0.23
GD-SD	6.57 ± 0.29	7.43 ± 0.15

between original and de-identified utterances. This means that the resulting speech sounds like a speaker that is neither the source nor the target of the conversion function, so the proposed method would not jeopardize the privacy of the source and target speakers. The MCD in a speaker-dependent setting should be investigated.

Finally, the listening test to evaluate similarity resulted in most of the original/de-identified pairs of utterances being labelled as "Different, absolutely sure", specially for GI-SI transformations. Nevertheless, for the GD-SI strategy, some pairs were rated as "Same, not sure", which suggests that this de-identification strategy is not as successful at removing the identity of the speaker. Also, GAN GD-SD system had several pairs rated as "Different, not sure". These results are consistent with those displayed in Fig. 4, where the three aforementioned transforms were the ones that achieved the lowest EERs.

7. Depression level estimation experiments

Three data partitions were established in AVEC 2014 depression sub-challenge, namely training, development and testing. However, according to Lopez-Otero et al. (2015), some of the speakers appeared both in training and test partitions, which led to biased results. Hence, in these experiments, a leave-one-speaker-out strategy was implemented to assess depression level estimation. Since this experimental configuration uses all the available data for training and testing, there is not a remaining set of recordings for parameter tuning; hence, as in Lopez-Otero et al. (2017b), the top performing configuration of the free parameters of the system (number of Gaussians of the UBM and dimension of the i-vectors) is presented for each experiment. UBMs of 256, 512 and 1024 mixtures and i-vectors of dimension 100, 200, 400 and 600 were considered. It must be noted that only the SVR is trained using AVEC 2014 training data, the UBMs and **T** matrices are trained using the Biosecure database (Section 5.1.4).

7.1. Evaluation metrics

Depression level estimation performance was measured by means of the root mean square error (RMSE) between the estimated and the groundtruth BDI of the recordings as defined in the experimental protocol of AVEC 2014 (Valstar et al., 2014). The mean absolute error (MAE) is also shown in order to compare it with the RMSE since the greater the difference between them, the greater the variance of the error.

7.2. Results

Table 4 shows the depression level estimation results achieved with original and de-identified speech based on FW+AS and GAN techniques. Results for gender-dependent and gender-independent approaches for depression estimation are presented in the table, as well as the achieved results per gender. Before discussing these results, it must be mentioned that the actual gender of the speaker was considered when performing gender-dependent depression level estimation; this means that, given a male speaker, depression estimation models trained on male voices are used regardless the apparent gender of the speaker after de-identification.

First, the validity of the depression level estimation strategy proposed in Section 4 must be discussed. As shown in Table 4, the RMSE of the estimated BDI levels is 9.82 when dealing with original speech in a gender-independent setting. Compared with Morales and Levitan (2016), where a leave-one-out strategy was also used to evaluate depression level estimation on AVEC 2014 recordings, it can be seen that the result using the i-vector based system exhibits a slightly better performance, which suggests that this approach is suitable for the analysis proposed in this paper. This was also validated in previous work by the authors (Lopez-Otero et al., 2017b).

The gender-independent depression level estimation results presented in Table 4 show that there are three de-identification strategies (FW+AS GI-SI, GAN GI-SD, GAN GD-SD) that lead to a performance that is significantly worse than that obtained with original speech. In addition, the performance of gender-dependent de-identification strategies show lower RMSE values than gender-independent strategies, except for GAN GD-SD.

The gender-dependent depression level estimation results show that, in general, training independent models for each gender leads to better estimation of the depression level. The difference is not so remarkable for original speech, but the difference is statistically significant when dealing with de-identified speech obtained with FW+AS, as shown in Table 4. However, for GAN-based de-identification, only the GI-SD de-identification strategy showed a significantly better performance when using

Table 4
Depression level estimation results using the proposed de-identification strategies. Results using the original recordings (i.e. without applying de-identification) are shown for comparison. Values of RMSE with superindex † show a statistically significant difference between original and de-identified results; values with superindex ‡ show a statistically significant difference between gender-independent and gender-dependent depression detection systems; values with superindex * show a statistically significant difference between FW+as and GAN systems. Two systems were considered to have a statistically significant difference when indicated by a paired *t-test* with a 95% confidence level.

		Gender ind.		Gender dep.		Male		Female	
Transform	Experiment	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
	Original	9.82	7.54	9.29	7.60	9.66	8.08	9.10	7.37
FW+AS	GI-SI	11.13†	9.81	8.78‡	7.08	9.06*	7.24	8.64	7.00
	GI-SD	11.08	8.21	9.24‡	7.09	9.97	7.83	8.87	6.73
	GD-SI	10.87	8.30	9.39‡	7.47	9.21	7.55	9.48	7.44
	GD-SD	10.70	8.33	8.93‡	7.15	10.07	8.51	8.33	6.49
GAN	GI-SI	10.82	8.09	10.18	7.71	11.12	8.96	9.70	7.10
	GI-SD	11.92†	9.41	9.16‡	7.23	9.48	7.44	9.00	7.13
	GD-SI	10.67	8.44	9.68	7.63	10.14	8.19	9.45	7.36
	GD-SD	11.55†	8.85	10.90	8.44	12.73	9.74	9.89	7.81

speaker-dependent models for depression estimation. Results for male and female speakers are also displayed in the table and they show that, in general, all the de-identification strategies achieved lower RMSE values for female speakers, except for the FW+AS GD-SI strategy.

Table 4 allows a comparison between GAN and FW+AS approaches when performing depression estimation. The table shows that, even though the difference in performance is non statistically significant for most of the transforms, the RMSE achieved by FW+AS is usually lower than its equivalent using the GAN conversion strategy.

As mentioned above, the actual gender of the speaker was considered when performing gender-dependent depression level estimation. Nevertheless, there is not guarantee that the depression estimation system has this information, so the most natural way to perform this experiment would be using the apparent gender of the speaker (i.e. the gender the speaker presents after de-identification). Table 5 shows a comparison of gender-dependent depression estimation when using the actual and apparent gender of the speaker. Only GI-SI and GI-SD de-identification strategies are displayed since GD-SI and GD-SD preserve the actual gender of the speaker hence the result would be the same in both cases. As shown in Table 5, the performance of the gender-dependent depression level estimation approach strongly decays when considering the apparent gender of the speaker. This means that, even though a de-identified male speaker sounds like a female, using male models for depression estimation is more suitable since these capture information that is specific for that gender (Cummins et al., 2017).

8. Discussion

The experimental results shown in Sections 6 and 7 highlight several interesting facts. First, it can be observed that GAN-based de-identification achieves higher EER values than FW+AS (i.e. the de-identified speakers are more different to the original speakers with GAN). However, this improvement in de-identification accuracy is reflected in the depression estimation experiments, where GAN-based transforms lead to a higher RMSE: the difference in RMSE is not significant according to a paired *t*-test with a 95% confidence level but, given the small size of the population (120 sessions), these results arise some concerns about the validity of GAN-based conversion functions for de-identification in the context of depression level estimation.

The use of speaker-dependent versus speaker-independent de-identification was also evaluated in this paper. The results show that both approaches lead to similar performance both in de-identification and depression estimation experiments. Given that speaker-dependent de-identification requires training data from the speakers to be de-identified, which is not usually available in a real setting, it can be concluded that the speaker-independent de-identification techniques evaluated in this paper are the best choice for real-world applications. The effort of compiling a parallel corpus for the users would be justified if the speaker-dependent performance was clearly superior than the speaker-independent one, but the results in this paper support the use of speaker-independent approaches, which are easier to apply in real scenarios.

Table 5
Depression level estimation results when considering the actual and apparent gender of the de-identified speech.

		Actual gender		Apparent gender	
Transform	Technique	RMSE	MAE	RMSE	MAE
FW+AS	GI-SI	8.78	7.08	12.07	9.88
	GI-SD	9.24	7.09	11.16	8.54
GAN	GI-SI	10.18	7.71	11.06	8.36
	GI-SD	9.16	7.23	15.10	12.45

Performance of depression level estimation when using gender-independent and gender-dependent models was also evaluated. The results showed that, in general, depression estimation using gender-dependent models lead to lower values of RMSE, but only when considering the original gender of the speaker (i.e. the gender of the spoken utterance before de-identification). However, when considering the apparent gender (i.e. the gender of the spoken utterance after de-identification), depression estimation results suffer a dramatic degradation. Hence, when performing gender-independent de-identification, using gender-independent depression level estimation is more reliable; and when performing gender-dependent de-identification, gender-dependent depression level estimation is the best alternative.

Besides automatic evaluation, it was mentioned in Section 1 that the quality and naturalness of the resulting speech is good enough so it can be used for depression evaluation by professionals or for research purposes. Listening to the de-identified recordings it is clear that the quality of the FW+AS recordings is much better than that achieved with GAN. This is expected based on MOS experiments performed in the literature; to cite some examples, quality MOS results presented in Magariños et al. (2017) are higher than GAN results with parallel training data presented in Fang et al. (2018), but still the performance of GAN is acceptable in terms of quality.

9. Conclusions and future work

This paper has presented an analysis of gender and identity issues in the context of depression level estimation of de-identified speech. The impact of performing gender-dependent or gender-independent speaker de-identification was assessed, and the results indicated that, even though the de-identification rate is lower for gender-dependent de-identification, it still succeeds at removing the identity of the speaker while achieving, in general, better depression level estimation results. The use of speaker-dependent and speaker-independent conversion functions for de-identification was also evaluated and, looking at the results, it can be concluded that both strategies achieve very similar performance in terms of de-identification rate and depression level estimation, so compiling speaker-specific corpora to train speaker-dependent conversion functions is not worth the effort. Despite this fact, the obtained results for speaker-dependent de-identification were interesting: it was demonstrated that it is possible to train FW+AS conversion functions using source speech whose conditions are not ideal for this task because of mispronunciations or noisy recordings, among other factors.

Listening tests were conducted by the authors of the paper since the EULA of AVEC data did not allow to share clips with non-signers of this document. The results of the similarity test showed a significant correlation with those achieved using a speaker verification system. In the future, it would be interesting to increase the number of participants in the listening test to validate this conclusion, since the possibility of evaluating speaker de-identification techniques using automatic tools would dramatically ease the research in this field. Another informal listening test, which was conducted to evaluate the quality of the de-identified speech, suggests that the quality of the resulting de-identified speech is good but still has room for improvement, specially in the case of the GAN-based transforms. Therefore, in future work, the use of other vocoders for voice conversion, such as Wavenet, will be investigated. In addition, other deep learning architectures besides GAN will be assessed and compared to the results presented in this work. Also, the assessment of some of the novel voice conversion approaches for non-parallel data (Saito et al., 2018a; Kaneko and Kameoka, 2018; Zhang et al., 2020) will be done in future work.

Given that, in this work, parallel corpora was compiled in which the depression level of the speakers is known, it would be interesting to train speaker de-identification and depression level estimation models using multitask learning techniques: this may lead to voice conversion functions that better preserve the depressive state of the patient while removing their identity from the speech recordings.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowlgedgments

This work has received financial support from i) “Ministerio de Economía y Competitividad” of the Government of Spain and the European Regional Development Fund (ERDF) under the research projects RTI2018-093336-B-C22 and RTI2018-101372-B-I00, ii) Xunta de Galicia (projects GPC ED431B 2019/003 and GPC ED431B 2018/60), and iii) Xunta de Galicia - “Consellería de Cultura, Educación e Ordenación Universitaria” and the ERDF through the 2016−2019 accreditations ED431G/01 (“Centro singular de investigación de Galicia”) and ED431G/04 (“Agrupación estratéxica consolidada”).

References

Abou-Zleikha, M., Tan, Z.-H., Christensen, M., Jensen, S., 2015. A discriminative approach for speaker selection in speaker de-identification systems. In: Proceedings of 23rd European signal processing conference (EUSIPCO), pp. 2147–2151.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Hyett, M., Parker, G., Breakspear, M., 2017. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. IEEE Trans. Affect. Comput.

Alhanai, T., Ghassemi, M., Glass, J., 2018. Detecting depression with audio/text sequence modeling of interviews. Interspeech, pp. 1716–1720.

Bahmaninezhad, F., Zhang, C., Hansen, J., 2018. Convolutional neural network based speaker de-identification. Odyssey 2018 The Speaker and Language Recognition Workshop, pp. 255–260.
Basu, S., 2003. A linked-HMM model for robust voicing and speech detection. In: Proceedings of ICASSP, vol. 1, pp. 816–819.
Beck, A., Steer, R., Ball, R., Ranieri, W., 1996. Comparison of beck depression inventories -IA and -II in psychiatric outpatients. J. Pers. Assess. 67 (3).
Ben-Zeev, D., Scherer, E., Wang, R., Xie, H., Campbell, A., 2015. Next-generation psychiatric assessment: using smartphone sensores to monitor behavior and mental health. Psychiatr. Rehabil. J. 38 (3), 218–226.
Chen, L., Lee, H.-Y., Tsao, Y., 2019. Generative adversarial networks for unpaired voice transformation on impaired speech. Interspeech, pp. 719–723.
Cohn, J., Kruez, T., Matthews, I., Yang, Y., Nguyen, M., Tejera Padilla, M., Zhou, F., De la Torre, F., 2009. Detecting depression from facial actions and vocal prosody. 3rd International Conference on Affective Computing and Intelligent Interaction, pp. 1–7.
Correia, J., Raj, B., Trancoso, I., 2018. Querying depression vlogs. In: Proc. of Spoken Language Technology Workshop (SLT), pp. 987–993.
Correia, J., Trancoso, I., Raj, B., 2016. Detecting psychological distress in adults through transcriptions of clinical interviews. Lect Notes Artif Intell 10077, 162–171.
Cummins, N., 2016. Automatic Assessment of Depression from Speech: Paralinguistic Analysis, Modelling and Machine Learning. The University of New South Wales Ph.D. thesis.
Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T., 2015. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 71, 10–49.
Cummins, N., Sethu, V., Epps, J., Schnieder, S., Krajewski, J., 2015. Analysis of acoustic space variability in speech affected by depression. Speech Commun. 75, 27–49.
Cummins, N., Vlasenko, B., Sagha, H., Schuller, B., 2017. Enhancing speech-based depression detection through gender dependent vowel-level formant features. Artificial Intelligence in Medicine (AIME 2017). Lecture Notes in Computer Science, vol. 10259. Springer, pp. 209–214.
Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2010. Front end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process.
Erro, D., Alonso, A., Serrano, L., Navas, E., Hernaez, I., 2015. Interpretable parametric voice conversion functions based on gaussian mixture models and constrained transformations. Comput. Speech. Lang. 30 (1), 3–15.
Erro, D., Moreno, A., Bonafonte, A., 2010. Voice conversion based on weighted frequency warping. Comput. Speech. Lang. 18 (5), 922–931.
Erro, D., Navas, E., Hernez, I., 2013. Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. IEEE Trans. Audio Speech Lang. Process. 21 (3), 556–566.
Erro, D., Sainz, I., Navas, E., Hernaez, I., 2011. Improved HNM-based vocoder for statistical synthesizers. Interspeech, pp. 1809–1812.
Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., Truong, K., 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. 7 (2), 190–202.
Fairbanks, G., 1960. Voice and Articulation Drillbook, second ed. Harper & Row.
Fang, F., Yamagishi, J., Echizen, I., Lorenzo-Trueba, J., 2018. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP.
Gao, S., Wu, X., Xiang, C., Huang, D., 2019. Development of a computationally efficient voice conversion system on mobile phones. APSIPA Trans. Signal Inf.Process. 8, E4.
Garcia-Romero, D., Espy-Wilson, C., 2011. Analysis of i-vector length normalization in speaker recognition systems. In: Proceedings of Interspeech, pp. 249–252.
Garfinkel, S., 2015. De-Identification of Personally Identifiable Information. Technical Report. National institute of standards and Technology (NIST), U.S. Department of Commerce.
Gravenhorst, F., Muaremi, A., Bardram, J., Grünerbl, A., Mayora, O., Wurzer, G., Frost, M., Osmani, V., Arnrich, B., Lukowicz, P., Tröster, G., 2015. Mobile phones as medical devices in mental disorder treatment: an overview. Pers. Ubiquitous Comput. 19 (2), 335–353.
He, L., Cao, C., 2018. Automated depression analysis using convolutional neural networks from speech. J. Biomed. Inform. 83, 103–111.
Hor, K., Taylor, M., 2010. Suicide and schizophrenia: a systematic review of rates and risk factors. J. Psychopharmacol. 24 (4), 81–90.
Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., Wang, H.-M., 2016. Voice conversion from non-parallel corpora using variational auto-encoder. 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA).
Huang, Z., Stasak, B., Dang, T., Wataraka Gamage, K., Le, P., Sethu, V., Epps, J., 2016. Staircase regression in OA RVM, data selection and gender dependency in AVEC 2016. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 19–26.
Jin, Q., Toth, A.R., Schultz, T., Black, A.W., 2009. Speaker de-identification via voice transformation. IEEE workshop on Automatic Speech Recognition and Understanding, pp. 529–533.
Justin, T., Štruc, V., Dobrišek, S., Vesnicer, B., Ipšic, I., Mihelič, F., 2015. Speaker de-identification using diphone recognition and speech synthesis. In: Conference and Workshops on Automatic Face Gesture Recognition, pp. 1–7.
Kaneko, T., Kameoka, H., 2018. CycleGAN-VC: non-parallel voice conversion using cycle-consistent adversarial networks. 26th European Signal Processing Conference, EUSIPCO, pp. 2100–2104.
Karam, Z.N., Provost, E., Singh, S., Montgomery, J., Archer, C., Harrington, G., Mcinnis, M., 2014. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 4858–4862.
Kipli, K., Kouzani, A., 2013. An algorithm for determination of rank and degree of contribution of sMRI volumetric features in depression detection. 35th Annual International Conference of the IEEE EMBS, pp. 1382–1385.
Kipli, K., Kouzani, A., Xiang, Y., 2013. An empirical comparison of classification algorithms for diagnosis of depression from brain sMRI scans. International Conference on Advanced Computer Science Applications and Technologies, pp. 333–336.
Lopez-Otero, P., Docio-Fernandez, L., Abad, A., Garcia-Mateo, C., 2017. Depression detection using automatic transcriptions of de-identified speech. In: Proceedings of Interspeech, pp. 3157–3161.
Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C., 2014. A study of acoustic features for depression detection. In: Proceedings of IWBF, pp. 1–6.
Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C., 2014. A study of acoustic features for the classification of depressed speech. In: Proceedings of MIPRO, pp. 1331–1335.
Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C., 2015. Assessing speaker independence on a speech-based depression level estimation system. Pattern Recognit. Lett. 68, 343–350.
Lopez-Otero, P., Magariños, C., Docio-Fernandez, L., Rodriguez-Banga, E., Erro, D., Garcia-Mateo, C., 2017. Influence of speaker de-identification in depression detection. IET Signal Proc.
Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., Ling, Z., 2018. The voice conversion challenge 2018: promoting development of parallel and nonparallel methods. Odyssey 2018 The Speaker and Language Recognition Workshop, pp. 195–202.
Magariños, C., Lopez-Otero, P., Docio-Fernandez, L., Banga, E., Erro, D., Garcia-Mateo, C., 2017. Reversible speaker de-identification using pre-trained transformation functions. Comput. Speech. Lang.
Magariños, C., Lopez-Otero, P., Docio-Fernandez, L., Erro, D., Banga, E., Garcia-Mateo, C., 2016. Piecewise linear definition of transformation functions for speaker de-identification. In: Proceedings of First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE), pp. 1–5.
Marcus, M., Yasamy, M., van Ommeren, M., Chisholm, D., Saxena, S., 2012. Depression: A Global Public Health Concern. Technical Report. World Health Organization.
Mitra, V., Shriberg, E., McLaren, M., Kathol, A., Richey, C., Vergyri, D., Graciarena, M., 2014. The SRI AVEC-2014 evaluation system. In: Proceedings of AVEC'14, pp. 93–101.
Mohammadi, S., Kain, A., 2017. An overview of voice conversion systems. Speech Commun. 88, 65–82.
Morales, M., Levitan, R., 2016. Speech vs. text: a comparative analysis of features for depression detection systems. IEEE Spoken Language Technology Workshop (SLT), pp. 136–143.
Morise, M., Yokomori, F., Ozawa, K., 2016. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans. Inf. Syst. E99-D (7), 1877–1884.

Nasir, M., Jati, A., Shivakumar, P., Chakravarthula, S., Georgiou, P., 2016. Multimodal and multiresolution depression detection from speech and facial landmark features. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 43–50.
Nolen-Hoeksema, S., 1987. Sex differences in unipolar depression: evidence and theory. Psychol Bull. 101 (2), 259–282.
Ortega-Garcia, J., Fierrez, J., Alonso-Fernandez, F., Galbally, J., Freire, M.R., Gonzalez-Rodriguez, J., Garcia-Mateo, C., Alba-Castro, J., Gonzalez-Agulla, E., Otero-Muras, E., Garcia-Salicetti, S., Allano, L., Ly-Van, B., Dorizzi, B., Kittler, J., Bourlai, T., Pho, N., Deravi, F., Ng, M., Fairhurst, M., Hennebert, J., Humm, A., Tistarelli, M., Brodo, L., Richiardi, J., Drygajlo, A., Ganster, H., Sukno, F., Pavani, S.-K., Frangi, A., 2010. The multiscenario multienvironment BioSecure multimodal database (BMDB). IEEE Trans. Pattern Anal. Mach. Intell. 32 (6), 1097–1111.
Pampouchidou, A., Simantiraki, O., Fazlollahi, A., 2016. Depression assessment by fusing high and low level features from audio, video, and text. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 27–34.
Paul, D., Pantazis, Y., Stylianou, Y., 2019. Non-parallel voice conversion using weighted generative adversarial networks. Interspeech, pp. 659–663.
Pitz, M., Ney, H., 2005. Vocal tract normalization equals linear transformation in cepstral space. IEEE Trans. Speech Audio Process. 13, 930–944.
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The Kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society.
Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted Gaussian mixture models. Digit. Signal Process. 10, 19–41.
Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., Pantic, M., 2017. AVEC 2017 - real-life depression, and affect recognition workshop and challenge. In: Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge (AVEC).
Sadjadi, S., Kheyrkhah, T., Tong, A., Greenberg, C., Singer, E., Mason, L., Hernandez-Cordero, J., 2017. The 2016 NIST speaker recognition evaluation. Interspeech, pp. 1353–1357.
Saito, Y., Ijima, Y., Nishida, K., Takamichi, S., 2018. Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors. ICASSP, pp. 5274–5278.
Saito, Y., Takamichi, S., Saruwatari, H., 2018. Statistical parametric speech synthesis incorporating generative adversarial networks. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (1), 84–96.
Sisman, B., Zhang, M., Sakti, S., Li, H., Nakamura, S., 2018. Adaptive WaveNet vocoder for residual compensation in GAN-based voice conversion. 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 282–289.
Sturim, D., Torres-Carrasquillo, P., Quatieri, T., Malyska, N., Mccree, A., 2011. Automatic detection of depression in speech using Gaussian mixture modeling with factor analysis. In: Proceedings of Interspeech, pp. 338–342.
Sun, B., Zhang, Y., He, J., Yu, L., Xu, Q., Li, D., Wang, Z., 2017. A random forest regression method with selected-text feature for depression assessment. In: Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 61–68.
Sündermann, D., Ney, H., 2003. VTLN-based voice conversion. In: Proceedings of ISSPIT, pp. 556–559.
Syed, Z., Sidorov, K., Marshall, D., 2017. Depression severity prediction based on biomarkers of psychomotor retardation. In: Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 37–43.
Valbret, H., Moulines, E., Tubach, J., 1992. Voice transformation using PSOLA technique. Speech Commun. 11 (2), 175–187.
Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M., 2016. AVEC 2016: depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 3–10.
Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., Pantic, M., 2014. AVEC 2014 3D dimensional affect and depression recognition challenge. In: Proceedings of AVEC'14.
Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M., 2013. AVEC 2013 - the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd International Worskhop on Audio/Visual Emotion Challenge, Proceedings of AVEC'13.
Veaux, C., Yamagishi, J., MacDonald, K., 2016. CSTR VCTK corpus: english multi-speaker corpus for CSTR voice cloning toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR).
Wester, M., Wu, Z., Yamagishi, J., 2016. Analysis of the voice conversion challenge 2016 evaluation results. Interspeech.
Williamson, J., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., Kung, H.-T., Dagli, C., Quatieri, T., 2016. Detecting depression using vocal, facial and semantic communication cues. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 3–10.
Williamson, J., Quatieri, T., Helfer, B., Ciccarelli, G., Mehta, D., 2014. Vocal and facial biomarkers of depression based on motor incoordination and timing. In: Proceedings of AVEC'14.
Yang, L., Jiang, D., He, L., Pei, E., Oveneke, M., Sahli, H., 2016. Decision tree based depression classification from audio, video and language information. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 43–50.
Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M., Sahli, H., 2017. Multimodal measurement of depression using deep learning models. In: Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 53–59.
Zhang, J., Ling, Z., Dai, L., 2020. Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 540–552.
Zhao, S., Nguyen, T., Wang, H., Ma, B., 2019. Fast learning for non-parallel many-to-many voice conversion with residual star generative adversarial networks. Interspeech, pp. 689–693.
Zhao, Y., Kuruvilla-Dugdale, M., Song, M., 2019. Voice conversion for persons with amyotrophic lateral sclerosis. IEEE J. Biomed. Health Inform.
Zorila, T., Erro, D., Hernaez, I., 2012. Improving the quality of standard GMM-based voice conversion systems by considering physically motivated linear transformations. Commun. Comput. Inf. Sci. 328, 30–39.(ISSN: 1865-0929)