Speaker De-identification using Diphone Recognition and Speech Synthesis

Tadej Justin[1], Vitomir Štruc[1], Simon Dobrišek[1], Boštjan Vesnicer[2], Ivo Ipšić[3] and France Mihelič[1]

[1] Faculty of Electrical Engineering, University of Ljubljana, Slovenia
[2] Alpineon Ltd, Ljubljana, Slovenia
[3] Department of informatics, University of Rijeka, Rieka, Croatia

***Abstract*— The paper addresses the problem of speaker (or voice) de-identification by presenting a novel approach for concealing the identity of speakers in their speech. The proposed technique first recognizes the input speech with a diphone recognition system and then transforms the obtained phonetic transcription into the speech of another speaker with a speech synthesis system. Due to the fact that a Diphone RecOgnition step and a sPeech SYnthesis step are used during the de-identification, we refer to the developed technique as DROPSY. With this approach the acoustical models of the recognition and synthesis modules are completely independent from each other, which ensures the highest level of input speaker de-identification. The proposed DROPSY-based de-identification approach is language dependent, text independent and capable of running in real-time due to the relatively simple computing methods used. When designing speaker de-identification technology two requirements are typically imposed on the de-identification techniques: *i)* it should not be possible to establish the identity of the speakers based on the de-identified speech, and *ii)* the processed speech should still sound natural and be intelligible. This paper, therefore, implements the proposed DROPSY-based approach with two different speech synthesis techniques (i.e, with the HMM-based and the diphone TD-PSOLA-based technique). The obtained de-identified speech is evaluated for intelligibility and evaluated in speaker verification experiments with a state-of-the-art (i-vector/PLDA) speaker recognition system. The comparison of both speech synthesis modules integrated in the proposed method reveals that both can efficiently de-identify the input speakers while still producing intelligible speech.**

I. INTRODUCTION

With the technological advances made over the last decades, people are now able to receive legal or medical advice from the comfort of their homes, interact with others through chat rooms, social networks and video-conferencing applications or use (virtual) personal assistants such as Apple's Siri or Microsoft's Cortana. Using these modern-day services and applications often involves sharing sensitive information that can easily be abused if the identity of the user is divulged. It is, therefore, of paramount importance to develop technology capable of protecting one's personal privacy by concealing the identities of the individuals captured in the given type of data (e.g., video, speech or text), while still preserving the relevant information contained in the data [1]. The technology exhibiting the outlined properties is usually referred to as de-identification technology.

In this paper we address the problem of speaker (or voice) de-identification, where our goal is to conceal the speaker identity in speech recordings and to ensure that the de-identified speech is still intelligible. We present a novel (language-dependent) approach to speaker de-identification that is based on diphone recognition and speech synthesis (we will refer to our approach as DROPSY in the remainder[1]). With the proposed approach a speech utterance is first subjected to a diphone recognition module that produces a sequence of recognized diphones, which are then fed to the speech synthesis module that generates the final de-identified speech. The result of the outlined approach is speech belonging to the speaker whose data was used during training of the speech synthesis module. Since every speech utterance in the correct language used as input to the DROPSY-based approach is "converted" into the speech of the same speaker, it is effectively de-identified. Note that the usage of the speaker de-identification approach as presented in this paper is limited to applications where the reverse process - to obtain the speaker's real identity - is not required. In general the approach is applicable in scenarios where the users either want to conceal their identity or are reluctant to transmit their natural speech through the communication channel, e.g. data line, telephone, due to security or other similar considerations.

Our DROPSY-based approach is fundamentally different from other existing techniques to speaker de-identification, which commonly belong to one of the two following groups: *i)* the group of voice-degradation approaches, or *ii)* the group of voice-conversion approaches.

Techniques from the first group (e.g.,[2], [3]) typically try to degrade speech in one way or the other with the goal of affecting the speaker recognition performance. These techniques exhibit on-line capabilities, but the result is often speech that is relatively unnatural or even unintelligible, as emphasized in [4]. In contrast, the intelligibility of the speech produced with our approach is largely dependent on the performance of the diphone recognizer and the synthesis technique used. While it is not possible to obtain robust automatic word recognition based solely on a phoneme-recognition system (a phonetic typewriter), our evaluation of the automatically recognized phonemes suggests that errors made by the recognition module are mostly realized as substitutions between phonetically similar phonemes. By

[1] **D**iphone **R**ecogniti**O**n and s**P**eech **SY**nthesis - DROPSY

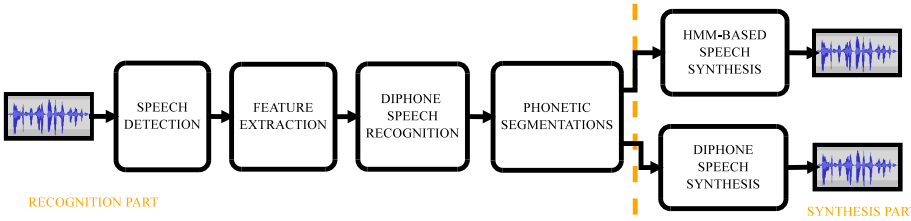

Fig. 1. Block diagram of the proposed DROPSY-based de-identification approach. The left part depicts the recognition module and the right part the synthesis module. Note that the right part of the figure shows the two synthesis techniques that are also used in our experimental evaluation in Section III.

listening to utterances with such substitution errors, the listener with the appropriate linguistic knowledge recognizes the uttered words and understands their meaning [5].

Techniques from the second group try to learn a mapping that transforms the voice of a source speaker to the voice of a target speaker. These techniques commonly require some speech material of the source speaker or, in some cases, even parallel corpora of aligned sentences of the source and target speakers to facilitate the successful estimation of the required mapping [6]. Once the mapping is learned the voice of the given input speaker can be converted to the voice of the target speaker and can consequently be de-identified. Our DROPSY-based approach, on the other hand, allows for the recognition and synthesis module to be trained separately without any speech material of the source speaker whose voice is to be de-identified.

Before we turn our attention to the description of the DROPSY-based approach in the next section, let us summarize the main contributions of the paper:

- We present a novel approach to speaker de-identification that relies on speech segment (diphone) recognition and synthesis of the recognized speech without the need for learning mapping functions between the source and target speakers.
- We implement the proposed approach using two speech synthesis techniques, i.e., the HMM-based and the diphone TD-PSOLA-based synthesis technique.
- We assess the developed de-identification approach in verification experiments with a state-of-the-art speaker verification system and show experimental results aimed at assessing the intelligibility of the de-identified speech.

The rest of the paper is structured as follows: In Section II we describe the proposed de-identification approach and all of its sub-parts. In Section III we present the verification experiments and intelligibility assessments and report about obtained results. Finally we conclude the paper in Section IV with some final comments and directions for future work.

II. DROPSY-BASED DE-IDENTIFICATION

The proposed DROPSY-based speaker de-identification approach is presented in Fig. 1 in the form of a block diagram. Our implementation consists of a bi-diphone speech recognition module, which performs the transformation of the input speech into the phonetic representation. The speech recognition module can also provide estimates of the pitch, loudness and duration, but since the output speech signal has to represent de-identified speech and these characteristics could provide clues towards the identity of the speaker, such estimates are discarded from further processing. With our approach the final output signal, i.e., the de-identified speech, is generated through synthesis of the phonetic representations, which results in a synthetic voice with the characteristics of the speaker., whose data was used during training of the speech synthesis module.

For our experiments, we implemented the DROPSY-based de-identification approach based on two speech synthesis techniques, i.e. the HMM-based synthesis technique and the diphone-based synthesis technique. While it is a generally acknowledged fact (see, e.g., [7]) that speech synthesized with the HMM-based approach is more natural than speech obtained through diphone-based synthesis, we included a comparison of intelligibility of both variants in our experiments due to the fact that diphone speech units were also the base units for our recognition module.

A. The recognition module

The speech recognition module is implemented using the Hidden Markov Model Toolkit (HTK) [10]. The speech recognition performance study presented in [9] has shown that the use of phone-transition-based speech units (such as diphones, bi-diphones, etc.) for acoustic modelling yields better speech recognition accuracy than the use of the traditional non-transition based phone models (such as monophones, biphones, triphones, etc.). The main differences between these two types of speech units are in the underlying speech segmentations, as the phone transition models do not represent the usual phone segments, but rather the transitions between the two "centres" of the subsequent phones. Since the concatenation points between the two subsequent phone-transition models are at the more stationary segments of the speech signal, the entropy of the speech decoding search process seems to be lower, and consequently the speech decoding accuracy is usually higher [9].

In our implementation of the speech (phone) recognition system the basic speech units are context-dependent diphones, called bi-diphones [8], that are modelled using the left-to-right continuous-density HMMs of three states with no state-skipping transitions, and with sixteen tied Gaussian mixtures per state. The usual MFCCs and energy plus the first and second order time derivatives are used as acoustic

features. A statistical phonetic-bigram language model is used to constrain the speech decoding search process.

B. The speech synthesis module

Two different speech synthesis modules are used in the experimental section for the implementation of our DROPSY-based de-identification approach. Each was trained with a different speech database; therefore, the synthesised voices from the two different implementations (i.e., the HMM- and the diphone-based implementations) have different target-speaker characteristics. In the first set-up the de-identified voice is obtained with the use of the diphone-based speech synthesis technique [12]. Similarly, in the second set-up, the de-identified speech is obtained with the speaker dependant HMM-based speech synthesis technique. The last was developed with the use of the HTS toolkit, version 2.2 [14], similar as in [13], where contextual quinphones were used for the base units.

C. Characteristics of the DROPSY-based approach

The proposed DROPSY-based de-identification approach exhibits the following characteristics:

- The de-identification approach is language dependent and text independent;
- The only requirement for the training data for the recognition and synthesis modules is that the training speech is uttered in the same language. There is no need to calculate any mapping from the source to the target speaker.
- The acoustical models of the synthesis module are completely independent of the acoustical models of the recognition module, which ensures the highest level of speaker de-identification (see Section III-C for empirical evidence).
- The synthesized de-identified speech produced by DROPSY is still intelligible (see Section III-A for details).

III. System evaluation setup and results

Note that two issues are important when assessing speaker (or voice) de-identification techniques: intelligibility of the de-identified speech and efficacy of the de-identification procedure. In the remainder of this section we present experiments aimed at evaluating both of these issues.

A. Intelligibility assessment

The proposed DROPSY-based de-identification approach is tested on the GOPOLIS speech database [15], which contains speech signals (read speech) and their transcriptions from 50 (25 male and 25 female) speakers. The word-recognition system using this database was developed and presented in [11]. The goal here (i.e, [11]) is to build an automatic speech dialogue system for querying flight information, thus, the vocabulary in the database is related to this task. Using the standard protocol defined for the database, the training set contains recordings from the first 18 male and 18 female speakers and the test set the recordings of the remaining 7 male and 7 female speakers. The training part of the database is used to train our bi-diphone speech recognition module.

For the evaluation of our speaker de-identification approach we randomly pick 28 test sentences from the test set with the following limitations:

- Only two sentences can be from the same speaker.
- Each sentence has to be between 5 and 8 words long.

With such limitations we ensure that the synthesized sentences are not too short, not too simple to understand, and on the other hand the sentences are not too long and consequently easy to forget. (remember that the evaluators task was to transcribe the recognized words from the artificial (de-identified) speech utterances). Using the outlined limitations we distribute the test sentences between all different test speakers. The final evaluation set consists of $2\times$(7 different male) and $2\times$(7 different female) input sentences, resulting in 56 (28 diphone synthesis, 28 HMM synthesis) different synthesized utterances.

All the evaluation tests were conducted with our own web-based evaluation system. Even though the application can be accessed from every computer with an internet connection and Adobe Flash enabled web browser, the evaluation process took place in a controlled environment at our faculty. During the evaluation process all the evaluators wore headphones. The evaluation test was successfully completed by 26 evaluators. All evaluators were 3rd year university-program students. The evaluators had a limited, basic knowledge of speech technologies.

Before the start of the evaluation task a brief description of the developed system and the evaluation process was given to the evaluators. During the introduction the semantic-domain (i.e., flight-service information queries) of the evaluation utterances was also explained. Each evaluator transcribed 7 sentences synthesized with the diphone speech synthesis system and 7 sentences synthesized with the HMM based synthesis system. The evaluators were divided into two groups. The first group with 13 evaluators evaluated the first randomly picked sentence from each speaker in the test dataset. The second group evaluated the sentences that were not evaluated by the first group of evaluators. With such an evaluation set-up we ensured that each evaluator listened to each sentence only once and also that all evaluators' transcriptions belonged to different input speakers. With our evaluation process we obtained a total of $(13+13)\cdot 7\cdot 2 = 346$ transcriptions.

The evaluated system for speaker de-identification can produce two kind of errors. The first one is related to the bi-diphone speech recognition module and can be measured in the form of the Phoneme Error Rate (PER). The second one can be presented as the output system error. This error represents a combination of influences from errors made by the speech recognition module, performance of the synthesis module and the evaluator capabilities. It can be measured from the analysis of evaluators' transcriptions in relation to the reference sentences transcriptions in the form of the Word Error Rate (WER).

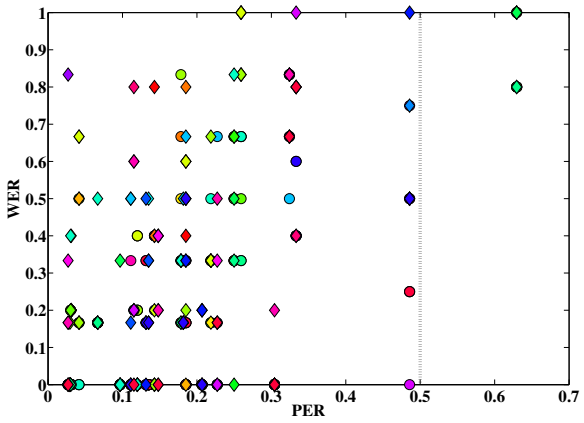

Fig. 2. Word error rate (WER) for all listening tests depending on the recognition phoneme error rate (PER). Points on the same vertical line match the transcriptions of different evaluators of the same test sentence. The color of the point shows the evaluator's identity. The evaluation was done for both speech synthesis modules - results for the HMM-based synthesis are marked as diamonds and for the diphone-based synthesis as circles.

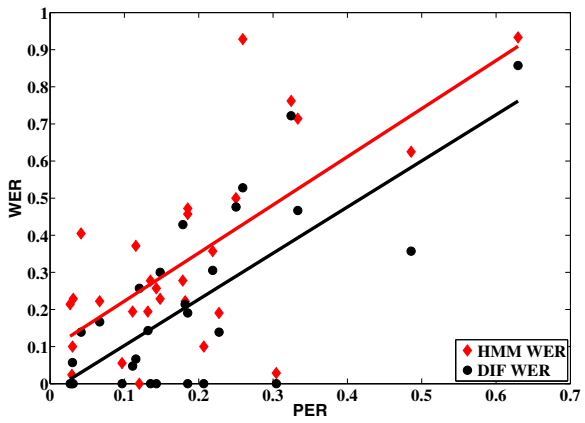

Fig. 3. Average word error rate (WER) for the HMM-based synthesis and diphone-based synthesis in relation to the phoneme error rate (PER). The HMM-based synthesis ($R^2 = 0.4806$) is shown in red and the diphone-based synthesis ($R^2 = 0.5382$) in black - a positive linear trend can be observed. Note that with the DROPSY-based approach the diphone-based synthesis is more intelligible than the HMM-based synthesis.

Note that the Error rate (ER) - either the Phoneme Error Rate or the Word Error Rate - is computed from the accuracy (A)

$$ER = 1 - A \quad ,$$

where the accuracy is determined as in [10] from the number of correctly recognized units N_{cor}, the number of deletion errors D and the number of reference units N_{ref}

$$A = \frac{N_{cor} - I}{N_{ref}} \quad .$$

B. Intelligibility evaluation results

The evaluation results presented in Fig. 2 can provide some interesting insights into the human ability to recognize complete words (and sentences) based on the acoustic representations produced in accordance with the results generated by a phoneme recognition system. As can be seen from the plots, humans are able to understand the majority of the words, despite the fact that phone-recognition errors occurred during the recognition step and that these errors were also propagated into the synthesized speech. This characteristic of human perception is also mimicked by automatic speech recognition systems, where it is commonly implemented with the help of a phonetic word lexicon, sentence-syntax and semantic-language modeling.

Each point in Fig. 2 represents the WER of the transcriptions produced by the evaluators in relation to the PER produced the recognition module. Points which are on the same vertical line match the transcriptions of the same test sentence, but are produced by a different evaluator. The color of the points shows the evaluator's identity, while the shape of the points indicates whether the HMM-based or the diphone-based speech synthesis is used. From Fig. 2 we can deduce that the results are strongly dependent on the identity of the evaluator. A detailed analysis shows that there are many transcriptions from the same evaluator that are recognized perfectly (WER = 0) or have a very low WER, although the bi-diphone recognition module never ensures a PER of zero (i.e., the phone error rate takes a non-zero value for all tested sentences).

Fig. 3 shows the trends of the average transcription WER (over the evaluators) for the input sentences in relation to the PER. As expected, we can observe a positive linear trend since lower values of the PER usually also result in lower values of the WER. Such an assumption can be verified by the significance test of the linear regression slope with the null hypothesis (B = 0). The significance test reveals that for both types of speech synthesis the null hypothesis is rejected with $p < 0,00005$.

We can also notice a difference between the average WER w.r.t. the type of speech synthesis used. The average WER for each type of speech synthesis is calculated from 182 test utterances and reveals the WER of 0.33 for the HMM-based approach and the WER of 0.21 for the diphone-based approach. These results are tested for significance with the binomial proportional test [16]. The test shows that the results are significantly different with the p-value of 0.005. The significance level is presented with the significance value at the 5% confidence interval. We can say that the de-identified speech synthesised with the diphone speech synthesis is significantly more intelligible than the de-identification speech synthesised with the HMM-based speech synthesis.

If we compare the obtained WER results with the WER from the word-recognition system evaluated for the task of a spoken dialogue system for air-flight queries [11], we notice - at first sight - a surprising paradox. The average WER 21 % obtained with the diphone speech synthesis is considerably higher than the WER 8% obtained in [11]. Thus, the machine speech recognition system outperforms the human recognition abilities? In this case this apparent paradox can be explained by the fact that the described word-recognition system used a relatively small word lexicon (829)

and a syntax model with a very low perplexity (5,7) [15].

Based on the evaluation results we can roughly estimate the value of the PER from input sentences where the evaluators could still understand the de-identified synthesized utterance. Certainly, we do not need a 0% recognition PER. In fact - as shown in Fig. 2 - in some occurrences the evaluators achieved the correct transcription (WER = 0), although the recognition PER was near 50%.

On the other hand, the PER and consequently the WER results are most likely also dependent on the input speaker's voice. For instance, we can see (Table I) that the PER and consequently the WER are dependent on the input speaker's gender. To statistically show the difference in the WER w.r.t the input speaker's gender, the results were tested for their significance with the binomial proportional test [16]. The HMM-based de-identified speech of a male input speaker is significantly more intelligible than the de-identified speech of a female input speaker, with a p-value of 0.001. If the input speaker was de-identified with the diphone-based speech synthesis module, male input speakers were significantly more intelligible than the female speakers with the p-value of 0.004. The significance level is presented with the significance value at the 5% confidence interval.

From these observations we can conclude that the DROPSY-based de-identification approach produces more intelligible speech for male input speakers than for female input speakers and also that the diphone-based speech synthesis module outperforms the HMM-based module in terms of intelligibility of the de-identified speech.

TABLE I

AVERAGE EVALUATORS WORD ERROR RATE (WER) FOR DIFFERENT TYPES OF SPEECH SYNTHESIS AND AVERAGE PHONEME ERROR RATE (PER) FOR ALL TEST UTTERANCES, DEPENDING ON SPEAKER GENDER.

Gender	WER HMM	WER DIF	PER
female	0,44	0,29	0,23
male	0,23	0,13	0,14

C. Speaker recognition evaluation

In our second series of experiments we try to evaluate the efficacy of the de-identification. To this end we implement an automatic state-of-the-art text-independent i-vector-based speaker recognition system [17]. The system used is a variant of the recognition system that ranked among the top 10 in the i-vector Machine learning Challenge organized as part of the 2014 Odyssey workshop in Finnland [21].

The system is trained on a subset of the NIST SRE 2004, 2005 and 2006 data, comprising telephone conversations of mostly English speech. In the acoustic front-end the system uses cepstral features extracted over 25 ms long overlapping windowed speech frames. Every 10 ms 19 Mel Frequency Cepstral Coefficients (MFCC) together with log-energy are calculated on the frequency range from 300 to 3400 Hz. Those 20 coefficients are augmented with their deltas and double deltas to produce the final 60-dimensional feature vector. The removal of non-speech frames is based on a simple energy-based-voice detector. For the extraction of Baum-Welch statistics we use the gender-independent Universal Background Model (UBM) consisting of 2048 diagonal Gaussians. The i-vector extractor produces 600-dimensional i-vectors. These i-vectors are projected to a 200-dimensional subspace with the Linear Discriminant Analysis (LDA) followed by length-normalization [18]. The final decisions scores are produced with the help of the Probabilistic Linear Discriminant Analysis (PLDA) classifier [19] consisting of 200 speaker- and 200 channel-factors.

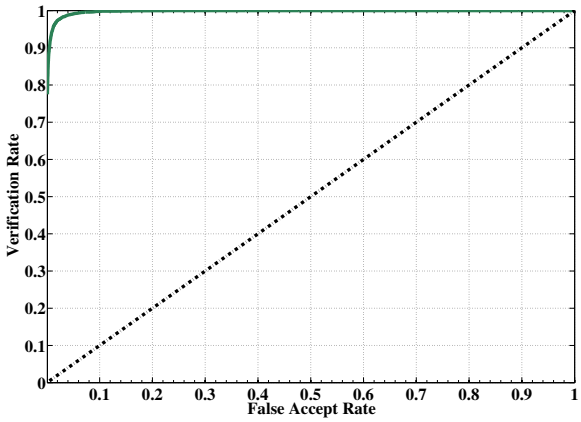

Fig. 4. The baseline performance of our speaker recognition system (used in verification experiments) in the form of a ROC curve (in green). The experiments were conducted with natural (not synthesized) speech. The dashed line indicated random performance.

For the purpose of the speaker recognition evaluation we conduct the evaluation test with the same test speaker identities, as were used in the intelligibility test. The target speakers are selected from the test sets of our database and were not used during training of the speech synthesis module or the diphone recognition module. To reduce the impact of speech-utterance durations on the speaker recognition performance we combine different test-utterances of the same speakers into speech samples of approximately 10-12 seconds. With this procedure we ensure that all speech utterances used in the experiments are of the same length. For each speaker we produce a total of 24 combined 10-12 seconds-long utterances. We produce ROC curves for all of our experiments.

D. Results of the speaker recognition evaluation

In the first experiment od this experimental series we establish the baseline performance of our speaker recognition system. The gallery consists of all available natural speech utterances (i.e. 384 utterances), including the natural speech of the speaker that was used to train the speech synthesis module. The test utterances also represent the natural speech of all the available speakers (i.e., 384 recordings). For this series of experiments 8832 legitimate verification attempts and 138240 illegitimate verification attempts are conducted.

It can be seen from Fig. 4 that the performance of the speaker recognition system on natural speech is reasonably

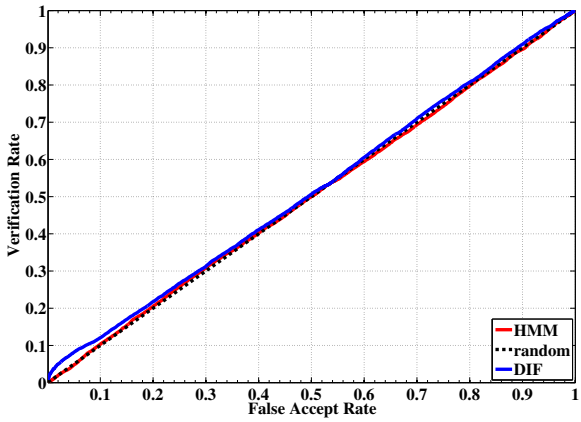

Fig. 5. The performance of the speaker verification test after the DROPSY-based de-identification with the HMM-based (in red) and diphone-based (in blue) speech synthesis module. The dashed line indicates random recognition performance. The graph is best viewed in color.

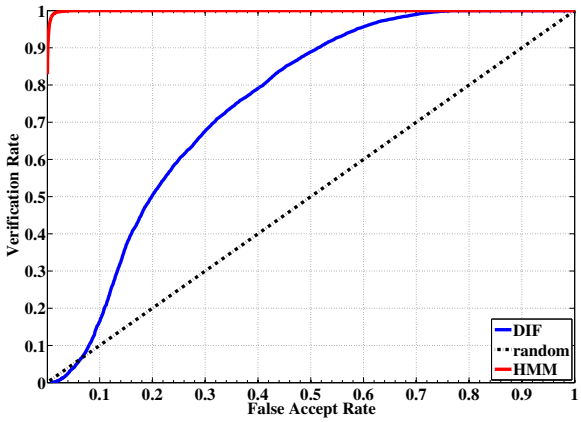

Fig. 6. The performance of the speaker verification test for the HMM-based (in red) and diphone-based (in blue) speech synthesis module. The test utterances represent de-identified speech, while the gallery represents natural (unaltered) speech recordings of various speakers including the speaker that was used to train the corresponding synthesis technique. Verification attempts between the target speaker and the de-identified speech are considered legitimate, while all other comparisons are treated as illegitimate when producing the ROC curves. The graph is best viewed in color.

high. On natural speech the system achieves a verification rate of 77.5% at 0.1% FAR and an Equal Error rate of 2.36%.

In the second experiment in this series we evaluate the efficiency of the DROPSY-based de-identification procedure. Here we try to evaluate whether it is possible to verify the identity of the speakers, when their speech is processed with the proposed de-identification procedure. Since we developed two independent speech synthesis modules with different target speaker characteristics, we repeat the experiment twice, the first time for the HMM-based synthesis and the second time for the diphone-based synthesis. In this experiment all speakers are again enrolled with the natural speech recordings, while the test data includes only recordings of de-identified speech. The identities of all speech recordings are left unaltered, i.e. if the original speech belongs to the subject *X*, we assume it belongs to the subject *X* after the de-identification as well. In Fig. 5 we present the results for the task of speaker verification - again in the form of ROC curves. To generate these ROC curves a total of 8280 legitimate and 129600 illegitimate verification attempts are conducted.

Note that the tested speaker recognition system is unable to recognize the true speaker identities from the de-identified speech with a performance better than chance. The ROC curves suggest that the recognition performance of both DROPSY-based implementations is more or less random. This result is also expected since all speech recordings were transformed to the speech of the speaker that was originally used to train the speech synthesis modules.

In our last experiment in this series we assess whether the speaker recognition system will indeed assign the de-identified speech to the speaker that was also used during the development of the speech synthesis modules. While the results of these experiments are not directly related to the efficiency of our de-identification procedure, they are nevertheless important as they have implications for other areas where DROPSY could be used, such as biometric spoofing where the goal is to compromise a biometric system by making some arbitrary speaker sound like a speaker known to the biometric system.

In this experiment the test utterances represent the de-identified version of our original test data, while the gallery represents recordings of natural speech including the two speakers that were used to train the HMM- and diphone-based synthesis modules. In experiments with the diphone-based synthesis module the test (probe, query) utterances represent only speech de-identified with the de-identification procedure with the diphone-based synthesis, and, similarly, in experiments with the HMM-based synthesis module the test (probe, query) utterances represent only speech de-identified with the de-identification procedure based on the HMM-based synthesis. For the ROC curve generation we re-label the test (de-identified) recordings and assume that all de-identified speech belongs to the target speaker that was used for training our corresponding synthesis module. Thus, if the de-identified speech is recognized as the speech of the target speaker, we should expect ROC curves with a high value of the AUC. This experiment includes 8640 legitimate and 17280 illegitimate verification attempts.

Fig. 6 shows that in the case of the HMM-based synthesis the verification performance is near optimal and the majority of the de-identified speech is correctly assigned to the target speaker. In the case of the diphone-based synthesis the result is not as convincing, but the de-identified speech is still assigned to the target speaker with the performance considerably different from chance.

IV. CONCLUSION

In this paper we proposed a novel method to the speaker-de-identification called DROPSY, which relies on a diphone-speech-recognition system and a speech-synthesis system to perform de-identification. We evaluated the proposed method through subjective listening tests to establish the intelligibil-

ity performance of the de-identified speech and also through the use of a speaker recognition system to assess the efficacy of the de-identification.

The proposed method can efficiently remove identity information from the input speech, while still producing speech that is intelligible in most cases. The reasons for such a behavior could be related to the performance of the speech recognition module. We showed that even though the recognition module does not ensure the PER of 0%, the de-identified speech can still be fully intelligible. The use of our DROPSY-based de-identification approach is a promising way of speaker de-identification since the results obtained on relatively small database suggest that it does not require a full-fledged and error-free speech recognition module. Nevertheless, there is still space for improvements and further experiments. One of the possible steps to achieve better results could be achieved with the use of the promising approach to speech recognition system, which is based on the deep-belief networks [22]. The main shortcoming of the proposed DROPSY-based de-identification is the naturalness of the de-identified speech, which will be the focus of our future efforts in this field. The proposed approach also lacks the ability to produce (de-identified) speech, from which it is possible to distinguish between different speakers, as all speech is mapped to the same target speaker. One of the possibilities to overcome this problem is to apply an acoustic transformation to the speech produced by the synthesis module, but would require performing some sort of speaker diarization on the input. Such s transformation can be easily applied when using the HMM-based speech synthesis module for producing the de-identified speech. Developing procedures for this next step will also be the part of our future work.

ACKNOWLEDGMENTS

The work presented in this paper was supported in parts by the national research program P2-0250(C) Metrology and Biometric Systems and the European Union's Seventh Framework Programme (FP7-SEC-2011.20.6) under grant agreement number 285582 (RESPECT). The support of COST Action IC1206 is also appreciated.

REFERENCES

[1] S. Ribarić et al., De-identification for privacy protection in mutlimedia content, *COST Action MOU*, 2013.

[2] Künzel, Hermann J. and Alexander, Paul Forensic Automatic Speaker Recognition with Degraded and Enhanced Speech,*J. Audio Eng. Soc*, vol. 62, num.4, 2014, pp 244-253

[3] Alegre, F.; Vipperla, R.; Evans, N.; Fauve, B., "On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals," *in Signal Processing Conference (EUSIPCO)*,Bucharest, Romania, 2012, pp. 36–40.

[4] Masthoff, Herbert. "A report on a voice disguise experiment. *International Journal of Speech Language and the Law*, vol 3, num 1, 2013, pp 160-167.

[5] Justin, Tadej and Mihelič, France and Dobrišek, Simon, Intelligibility Assessment of the De-Identified Speech Obtained Using Phoneme Recognition and Speech Synthesis Systems, *Lect. notes comput. sci.*, vol. 8655, 2014, pp. 529-536

[6] Stylianou, Y., "Voice Transformation: A survey",Formosa, Taiwan, in ICASSP-2009, 2009, pp. 3585-3588.

[7] Boštjan Vesnicer and France Mihelič, Evaluation of the Slovenian HMM-based speech synthesis system, *Lect. notes comput. sci.*, Vol 3206, 2004, pp. 513-520.

[8] Dobrišek, Simon and Mihelič, France and Pavešić, Nikola, "Acoustical modelling of phone transitions: biphones and diphones - what are the differences?", *in Sixth European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999, pp 1307–1310.

[9] S. Dobrišek, Analysis and Recognition of Phones in Speech Signals, *PhD Thesis, University of Ljubljana*, 2001.

[10] Young, S.J., Evermann, G., Gales, M.J.F., et al.: The HTK Book, version 3.4.1. Cambridge University Engineering Department, Cambridge, UK. (2009)

[11] I. Ipšić, F. Mihelič, S. Dobrišek, J. Gros, N. Pavešič, "A Slovenian spoken dialog system for air flight inquiries", *in 6th European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999, vol. 6, pp. 2659–2662.

[12] J. Žganec Gros, Jerneja, N. Pavešić, F. Mihelič, Text-to-Speech synthesis: a complete system for the Slovenian language', *CIT*, Vol. 5, nr. 1, pp. 11-19.

[13] M. Pobar, T. Justin, J. Žibert, F. Mihelič, I. Ipšić, A Comparison of Two Approaches to Bilingual HMM-Based Speech Synthesis, *Lect. notes comput. sci.*, Vol 8082, pp. 44-51.

[14] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W Black, K. Tokuda, "The hmm-based speech synthesis system (hts) version 2.0", *in Proc. of Sixth ISCA Workshop on Speech Synthesis*, Bonn, Germany, 2007, pp. 294-299 .

[15] F. Mihelič, J. Žganec Gros, S. Dobrišek, J. Žibert, N. Pavešić, Spoken language resources at LUKS of the University of Ljubljana, *Int. j. speech technol.*, vol. 6, iss. 3, 2003, pp. 221-232.

[16] R. B. D'agostino, W. Chase, and A. Belanger, "The ap propriateness of some common procedures for testing the equality of two independent binomial populations" , *The American Statistician*, vol. 42, no. 3, 1988, pp. 198–202.

[17] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P and Ouellet, P., Front-End Factor Analysis for Speaker Verification, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19(4), 2011, pp. 788-798.

[18] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems", *in Proceedings of Interspeech*, Florence, Italy, 2011, pp. 249-252.

[19] Kenny, P., "Bayesian Speaker Verification with Heavy-Tailed Priors", *in Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[20] Jin, Qin and Toth, Arthur R and Black, Alan W and Schultz, Tanja, "Is voice transformation a threat to speaker identification?.", *in ICASSP 2008*, Las Vegas,Nevada,2008, pp. 4845–4848

[21] Boštjan Vesnicer and Jerneja žganec-Gros and Simon Dobrišek and Vitomir Štruc, "Incorporating Duration Information into I-Vector-Based Speaker Recognition Systems", *in Odyssey: The Speaker and Language Recognition Workshop, Joensuu*, Finland, 2014.

[22] Graves, Alex and Mohamed, A-R and Hinton, Geoffrey, "Speech recognition with deep recurrent neural networks", *ICASSP 2013*,Vancouver, BC, Canada, 2013, pp. 6645–6649