Online speaker de-identification using voice transformation

M. Pobar, I. Ipšić

University of Rijeka/Department of Informatics, Rijeka, Croatia
{mpobar, ivoi}@inf.uniri.hr

***Abstract*—Speaker de-identification is the process by which speech is transformed in a way that the speaker identity is masked, while at the same time the transformed speech preserves acoustic information that contributes to the intelligibility, naturalness and clarity. Systems that perform speech de-identification could be used in voice driven applications (for example in call centres) where the speaker's identity has to be hidden.**

The paper describes the experiments we have performed in order to de-identify speech using GMM based voice transformation techniques and speaker identification using freely available tools.

We propose a method by which speakers whose speech has not been used to build voice transformations (for training) can be efficiently de-identified online.

The proposed method is evaluated using a speech database of read speech and a small set of speakers.

The results we present show that the proposed de-identification method performs similarly as a closed-set de-identification procedure that requires previous enrolment and can efficiently be used for online speaker de-identification.

***Index Terms*—speaker de-identification, voice transformation, online de-identification**

I. INTRODUCTION

In recent times many useful services have become available via the web or over the telephone. With the increased usage of such services, the users are also becoming more aware of privacy implications of their use. Therefore applications that can assure that users can protect their privacy are becoming more attractive. Methods concerning person de-identification in still images or video have already been proposed [1], and try to mask identification features such as faces, silhouettes, posture, gait etc.

There is also a need for de-identification technologies in voice driven applications. For example, conversations may be recorded in call centres for various purposes, such as analysis of operator mistakes, making the communication protocol more efficient in general or to prove a call has actually been made in case of complaints etc. In many cases, the identity of the caller is not important for the given purpose, and customers may legitimately wonder why it should be recorded and kept.

These concerns could be addressed by using a speaker de-identification process, where speech is altered in such a way that the identity of the speaker cannot be determined from the acoustic features, but speech itself remains intelligible. Voice transformation methods could be used for de-identification of speech, with appropriate selection of the target speaker. Voice transformation is used to make speech from one *source* speaker sound like it was uttered by another, *target* speaker.

In a commonly used GMM-based voice transformation scheme, speech from both source and target speakers has to be available to train the transformation, which can then be applied to novel speech from the source speaker and transformed into speech that sounds like the target speaker had uttered it. For the purpose of de-identification, it would be desirable that the target speaker is not a real person, as that could itself pose privacy problems for that person, but a synthetic surrogate speaker whose voice does not belong to any single individual.

In [6], [5] voice transformations were implemented to de-identify a small set of speakers and tested with automatic speaker identification systems. Voice transformation was successful in concealing identities of source speakers against the GMM-based speaker identification system, and a modified scheme was also successful with phonetic-based speaker identification (SID). However in those experiments, speech samples from each person to be de-identified had to be available in advance in order to estimate the transformation parameters. In addition, those samples were parallel utterances, with the same text spoken by source and target speakers. Also, to de-identify a speaker, his identity has to be known first, so that his corresponding voice transformation can be used for de-identification. This may also be a limitation in some cases, where the user doesn't want to identify with the system at all, such as in cases of anonymous police or help lines.

In a scenario with a closed set of speakers to be de-identified, this may be acceptable, but for applications like call centres etc., it would not be practical. In that case, the number of potential users of the system is extremely large, and many users will only use the system once. A requirement that the user has to supply a number of speech samples simply to use the system would be inconvenient. For a practical application in such systems, it would be desirable that any new user can use the system immediately, without having to enrol with the system first in any way or to identify himself, even for the purpose of de-identification.

In this paper, we propose a novel scheme for speaker de-identification where a set of pre-calculated voice transformations is used to de-identify new, unseen users' speech. Automatic speaker identification within this set is used to select the appropriate transform, which is then used to de-identify speech from the new user. We test the effectivnes of this method using automatic speaker identification and compare it with results of de-identification with previous enrolment.

The rest of the paper is organized as follows: In the next section we describe the speech database used in the

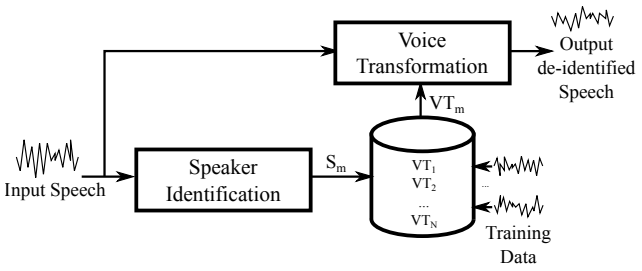

Figure 1. System block diagram.

Table I
SUMMARY OF DATA PER SPEAKER.

	Speaker				
	m01	m02	m03	m04	m05
Duration (m:s)	29:44	0:18	10:04	32:30	21:24
train. utts.	269	6	82	307	170
test utts.	66	1	20	76	42
	Speaker				
	m06	m07	m08	m09	m10
Duration (m:s)	1:26	2:22	34:40	28:40	24:49
train. utts.	21	45	293	253	213
test utts.	5	10	73	62	52

experiments and the setup of the speaker identification and voice transformation systems. Next we describe the performed de-identification experiments and present the results. Finally we give some conclusions and suggest future work.

II. EXPERIMENTAL SETUP AND RESULTS

A. Database description

In our experiments, we used a subset of the VEPRAD [8] database of spoken Croatian radio news containing both read and spontaneous speech. This is a multi-speaker database with male and female speakers, but we only used speech from 10 male speakers in this experiment, labelled m01-m10. 7 speakers have more than 20 minutes of speech, two have about 2 minutes and one speaker has only 20 seconds of data. Table 1 shows the statistics per speaker of used data. The speech is sampled with 16000 kHz and at 16 bit resolution.

For the synthetic voice, used as the target for voice transformation, we used 17 minutes of synthesized utterances from the same domain as the natural speech.

B. Voice transformation

The goal of voice transformation (VT) systems is to modify speech from one speaker (source speaker) so that it sounds like it was uttered by another (target) speaker. These systems learn a transformation function from speech data of source and target speakers. In most cases, the systems need a parallel corpus containing recordings of same sentences uttered by both source and target speakers, so that the recordings have the same phonetic content. The aligned speech data is used to calculate the transformation function that maps the source speaker's acoustic space into the target speaker space. The requirement for parallel corpus is too limiting for practical application in de-identification, as new users that need to be de-identified would first have to enrol in the system by providing the recordings of those specific sentences. In our experiment, we use a pool of pre-calculated transformations trained on data from an initial set of enrolled speakers, so we only need data for those speakers, which is not a demanding task, performed off line in the training phase. Speech from new users is de-identified using one of the transformations from this set, without additional training data from the new speaker. Avoiding the need for parallel corpus even for training of this initial set of transformations however additionally has an advantage, as during the system's usage data from new, unseen speakers can be used to train new transformation functions and thus expand the pool of available transformations. This could potentially improve the performance of the de-identification system. Data collected during the use of the system most certainly won't contain the same utterances as the target training corpus. With this in mind, we chose a voice transformation system that can train the transformation functions on non-parallel corpora.

We used a freely available voice transformation system [3], [4] based on GMM mapping and harmonic plus stochastic models (HSM) [9].

The target speaker was a synthetic HMM-based voice trained with the HTS [11] system. The target voice was trained using speech from four different male speakers from the VEPRAD database. Two of these speakers were also used for training the UBM for speaker identification, and are tested for de-identification, while speech from the other two speakers is only used for the average voice. Duration of training data for the target speaker is 17min 29s. Speaker adaptive training [10] was used to obtain the average voice. Normally, in speech synthesis this voice is then adapted to a desired speaker's identity, but for the purpose of de-identification, the "averageness" of the voice was actually desirable. Figure 1 depicts the architecture of the online de-identification system.

To train the transforms, audio data was first processed to obtain the HSM parameters, using the tools provided with the VT system. For this task pitch marks were required for each audio file, which were generated using the DYPSA [7] pitch extraction algorithm.

C. Speaker identification

We used the open-source Alize/SpkDet [2] platform for speaker recognition in our experiments. The system was used in two ways.

First, to verify the success of de-identification, so if the speaker identification system could correctly identify the original speaker from the speech transformed via VT, the de-identification would not be successful.

The same system was also used to select the most appropriate voice transform to use for de-identification, by choosing the transform belonging to the "most similar" speaker to the (unknown) speaker of a utterance to be de-identified. Speech from 10 male speakers was used to train 10 voice transformations between each speaker and the synthetic target speaker.

The speaker recognition system was trained to recognize the same 10 speakers, using the data specified in Table 1. When a sample of speech from a speaker who is unknown to the system has to be de-identifed, we first run it through the speaker recognition system. The system is forced to identify the speech as produced by one of the 10 known speakers, and we apply the transformation trained on that speaker to the speech sample. Due to limited data, we used the same speakers to train the voice transformations and to test the de-identification performance. For this reason, we actually used 9 possible transformations for each tested speaker, as we excluded that speaker's transformation from consideration. It is assumed that the speaker whose speech is to be de-identified sounds similar enough to the chosen speaker so that the transformation both de-identifies the speaker and keeps the intelligibility.

The system was used in a classical Gaussian mixture model (GMM) configuration.

The UBM was trained using data from 6 speakers (m02, m04, m05, m06, m07 and m10), with total duration of 1h 35 min. The individual GMMs were trained using 80% of data from each speaker. The same training set was used for training the GMMs for speaker identification and for training the voice transformations in the previous step, while the test set was used only in the de-identification tests.

We used filter-bank based cepstral features, with 24 filters in the filter bank. Along with 19 cepstral coefficients, log energy, and first and second order derivatives of features were used. Pre-emphasis coefficient was 0.97 and liftering value was 22. The UBM model is composed of 32 Gaussian components with diagonal covariance matrices.

D. Experiments and results

Each sample utterance from the test set was first fed into the speaker identification system unmodified, to determine the baseline performance of the speaker identification system. Out of 407 samples, 397 were correctly recognized, or 97.54%. Only those speech samples that were correctly recognized were used in subsequent de-identification experiments. The confusion matrix for the case of unmodified speech is shown in Table II. The speakers are labelled m01-m10, and the target speaker for de-identification is labelled *target*. The rows in the matrix represent the true speaker of an utterance, while the columns represent the hypothesized speaker as output from the speaker identification system. The values in each cell represent the number of test utterances.

In the first de-identification experiment, we transformed the voices of the 10 source speakers to the target synthetic speaker and performed the automatic speaker identification using the GMMs trained on natural speech from these speakers. To transform each source speaker, transformation trained with that speaker's data was used. This is similar to the scenario investigated in [6], where a closed set of speakers whose transforms had been trained from their own data can be de-identified. This case is our high baseline. Our assumption was that with the second approach that allows de-identification of unknown speakers we could achieve the same de-identification efficiency as with the baseline. For the baseline case, the SID system correctly identified the real speaker in 9% of cases after de-identification, giving de-identification rate of 91%. Further broken down, in 72% of cases when the original speaker is successfully de-identified, the detected speaker is the target (synthetic) speaker, and in 28% of the cases some other speaker from the closed set has the highest likelihood score. The full confusion matrix is given in Table III.

Table II
CONFUSION MATRIX, UNMODIFIED SPEECH.

True speaker	Hypothesized speaker										
	target	m01	m02	m03	m04	m05	m06	m07	m08	m09	m10
target	0	0	0	0	0	0	0	0	0	0	0
m01	0	62	0	1	1	0	0	0	0	0	2
m02	0	0	1	0	0	0	0	0	0	0	0
m03	0	0	0	20	0	0	0	0	0	0	0
m04	0	0	0	0	75	0	0	0	1	0	0
m05	0	0	0	0	0	41	0	1	0	0	0
m06	0	0	0	0	0	0	5	0	0	0	0
m07	0	0	0	1	0	0	1	6	1	0	1
m08	0	0	0	0	0	0	0	0	73	0	0
m09	0	0	0	0	0	0	0	0	0	62	0
m10	0	0	0	0	0	0	0	0	0	0	52

The second experiment concerns our proposed procedure, where de-identification of speech from a novel speaker is done with transforms trained on data from another speaker. The de-identification scheme consists of voice transformation system with a set of transformations trained from data of multiple speakers, and a speaker identification system trained to recognize the same speakers whose data was used to train the voice transformation parameters. When a novel speech sample from an unknown speaker is presented to the whole system, first the transformation to be applied is chosen based on speaker identification results. The speaker whose model has the highest log-likelihood ratio of having produced that sample is selected, and the transformation learned from that speaker's data is applied to the speech sample of the unknown speaker.

Due to a small number of speakers in the experiment, we used a leave-one-out scheme and used the data from the same speakers for training the transformations and for testing the de-identification. For each speech sample, the GMM corresponding to that speaker was excluded from the SID system, so that the speaker identification system could not identify the real speaker but had to choose from one of the remaining 9. The corresponding transform was applied to each sample to de-identify it. Then, all samples were fed again into the speaker identification system to verify the success of de-identification. For verification, all GMMs including the one for the speaker chosen for de-identification were used in the SID system. De-identification was successful in 87.41% of samples in this case. Out of all these de-identified samples, 70% were identified as the target (synthetic) speaker, and 30% as some other, wrong speaker, the same as in case of de-identification with previous enrolment. Full confusion matrix is shown in Table IV.

Table III
CONFUSION MATRIX, CLOSED-SET DE-IDENTIFICATION.

True speaker	Hypothesized speaker										
	target	m01	m02	m03	m04	m05	m06	m07	m08	m09	m10
target	0	0	0	0	0	0	0	0	0	0	0
m01	20	0	0	19	0	0	1	0	0	22	0
m02	1	0	0	0	0	0	0	0	0	0	0
m03	6	0	0	14	0	0	0	0	0	0	0
m04	44	0	0	12	0	1	0	7	0	11	0
m05	28	0	0	2	0	0	1	0	0	10	0
m06	5	0	0	0	0	0	0	0	0	0	0
m07	5	0	0	0	0	0	0	1	0	0	0
m08	63	0	0	5	0	0	3	2	0	0	0
m09	42	0	0	0	0	0	0	0	0	20	0
m10	47	0	0	0	0	0	1	0	0	4	0

Table IV
CONFUSION MATRIX, ONLINE DE-IDENTIFICATION.

True speaker	Hypothesized speaker										
	target	m01	m02	m03	m04	m05	m06	m07	m08	m09	m10
target	0	0	0	0	0	0	0	0	0	0	0
m01	19	1	0	18	0	0	0	0	0	24	0
m02	1	0	0	0	0	0	0	0	0	0	0
m03	5	0	0	15	0	0	0	0	0	0	0
m04	52	0	1	5	1	0	5	7	0	4	0
m05	32	0	0	2	0	0	1	0	0	6	0
m06	5	0	0	0	0	0	0	0	0	0	0
m07	4	0	0	0	0	0	1	1	0	0	0
m08	59	0	0	3	0	0	0	0	11	0	0
m09	41	0	0	0	0	0	0	0	0	21	0
m10	27	0	0	2	0	0	1	1	0	21	0

III. CONCLUSIONS AND FUTURE WORK

In the paper we examined a novel scheme for speaker de-identification built upon the idea of using voice transformation for de-identification. The method does not require enrolment of speakers for de-identification thus greatly extending possible applications of the system. The results show that the proposed method gives similar de-identification rate in comparison with previously available research, but with added flexibility. The system was tested using 10 speakers, and was able to successfully de-identify speakers in 87.4% of the tested cases. The performance is very close to the baseline de-identification system that requires previous enrolment of users, which was successful in fooling the speaker identification system in 91% of the tested cases.

In [6] some improvements to baseline voice transformation were proposed. These modifications should be tested within the proposed framework to examine if the performance of the new framework will scale similarly as in the closed set case.

The system can be extended to improve itself with usage if speech from new users using the system is used to train additional transforms, in effect expanding the available pool of transforms. That way, each new unseen speaker has a better chance of being more close acoustically to a known speaker and the corresponding transform may fit better to the purpose of de-identification.

Since promising de-identification results were obtained using the small test database, we plan to test the system with both more speakers (male and female) and more data per speaker. In that case, more data will be used to train the speaker identification system, which should make de-identification harder, while data for training the transformations will be varied to keep the scenario for online de-identification valid.

The naturalness and intelligibility of the de-identified speech depends on the quality of voice transformation and of the target speaker, which was in this case a synthetic average voice of several speakers. Listening to the de-identified speech we conclude that it is intelligible and has a certain amount of vocoded buzzy character that was also present in the target speaker's speech. However, a detailed formal subjective evaluation of de-identified speech with more listeners will also be performed to verify that the intelligibility of de-identified speech is not significantly degraded.

REFERENCES

[1] Prachi Agrawal and PJ Narayanan. Person de-identification in videos. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(3):299–310, 2011.

[2] Jean-François Bonastre, Nicolas Scheffer, Driss Matrouf, Corinne Fredouille, Anthony Larcher, Alexandre Preti, Gilles Pouchoulin, Nicholas Evans, Benoıt Fauve, and John Mason. Alize/spkdet: a state-of-the-art open source software for speaker recognition. *ISCA-IEEE Speaker Odyssey*, 2008.

[3] Daniel Erro and Asunción Moreno. Frame alignment method for cross-lingual voice conversion. *tC*, 2(1):1.

[4] Daniel Erro and Asunción Moreno. Weighted frequency warping for voice conversion. In *INTERSPEECH*, pages 1965–1968, 2007.

[5] Qin Jin, Arthur R Toth, Tanja Schultz, and Alan W Black. Speaker de-identification via voice transformation. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 529–533. IEEE, 2009.

[6] Qin Jin, Arthur R Toth, Tanja Schultz, and Alan W Black. Voice convergin: Speaker de-identification by voice transformation. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3909–3912. IEEE, 2009.

[7] A. Kounoudes, P. Naylor, and M. Brookes. The DYPSA algorithm for estimation of glottal closure instants in voiced speech. In *IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING*, volume 1. IEEE; 1999, 2002.

[8] S. Martincic-Ipsic and I. Ipsic. Veprad: a croatian speech database of weather forecasts. In *Information Technology Interfaces, 2003. ITI 2003. Proceedings of the 25th International Conference on*, pages 321–326, 2003.

[9] Ioannis Stylianou. *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, 1996.

[10] J. Yamagishi, T. Masuko, K. Tokuda, and T. Kobayashi. A training method for average voice model based on shared decision tree context clustering and speaker adaptive training. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 1, pages I–716–I–719 vol.1, 2003.

[11] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W Black, and Keiichi Tokuda. The hmm-based speech synthesis system (hts) version 2.0. In *Proc. of Sixth ISCA Workshop on Speech Synthesis*, pages 294–299, 2007.