Speaker De-identification via Voice Transformation

Qin Jin [#1], Arthur R. Toth [#2], Tanja Schultz [#3], Alan W Black [#4]

[#] *Language Technologies Institute, Carnegie Mellon University*
Pittsburgh, PA 15213, USA

[1] qjin@ca.cmu.edu
[2] atoth@cs.cmu.edu
[3] tanja@cs.cmu.edu
[4] awb@cs.cmu.edu

***Abstract*—It is a common feature of modern automated voice-driven applications and services to record and transmit a user's spoken request. At the same time, several domains and applications may require keeping the content of the user's request confidential and at the same time preserving the speaker's identity. This requires a technology that allows the speaker's voice to be de-identified in the sense that the voice sounds natural and intelligible but does not reveal the identity of the speaker. In this paper we investigate different voice transformation strategies on a large population of speakers to disguise the speakers' identities while preserving the intelligibility of the voices. We apply two automatic speaker identification approaches to verify the success of de-identification with voice transformation, a GMM-based and a Phonetic approach. The evaluation based on the automatic speaker identification systems verifies that the proposed voice transformation technique enables transmission of the content of the users' spoken requests while successfully preserving their identities. Also, the results indicate that different speakers still sound distinct after the transformation. Furthermore, we carried out a human listening test that proved the transformed speech to be both intelligible and securely de-identified, as it hid the identity of the speakers even to listeners who knew the speakers very well.**

I. INTRODUCTION

There are multiple aspects to the area of speaker identification. The most common is to use automatic methods to identify which speaker is speaking. But technology that makes it hard to identify a speaker also has its uses. We envision a number of scenarios where masking a speaker's voice is important. For example, in doctor-patient interviews, it may be necessary to mask the patient's voice such that the speech is still fully intelligible and remains natural sounding, but such that the listeners can no longer recognize the identity of the original speaker.

In earlier work [1] we proposed using voice transformation techniques to de-identify speech. In this paper we expand on that notion to further test our techniques on much larger populations of speakers. De-identification with larger groups of speakers could, theoretically, be easier. As the number of speakers increases, the probability that a modified speaker will be confused with another speaker increases. Therefore, we have used empirical tests to evaluate the different hypotheses. More importantly we also show through human listening tests that the de-identification process is very successful, and show that even after de-identification the generated voices can still be distinguished between each other even when they may not be related back to the original speaker. This property is important when conversations of multiple speakers are de-identified as we wish the resulting voices still to sound different from each other so that the listener is able to discriminate between the contributions of different speakers.

The technology of voice de-identification nicely complements other privacy and security related technologies, such as face de-identification [2] and word choice de-identification [3]. While our techniques perform de-identification only based on modifications of spectral and prosodic features, we are aware that other aspects of the de-identified speech, such as the speaker's word choice, may still hold residual identification information. However, we feel that de-identification technologies should offer choices of what level and aspects of de-identification should be applied.

The paper is organized as follows. Section II describes the voice transformation system and the four approaches we explored for speaker de-identification. Section III introduces the two speaker identification systems used for evaluating the success of de-identification. In section IV, we present the de-identification performance via voice transformation, the human listening test results, and distinguishability of de-identified voices. We conclude in Section V with a summary of our findings and suggest several avenues for future work.

II. VOICE TRANSFORMATION (VT)

Voice transformation (VT) attempts to make speech spoken by the source speaker sound as if it were produced by the target speaker. It can be applied to speaker de-identification since one strategy for de-identifying speech from multiple speakers is to transform it such that it sounds like it was all produced by the same speaker. In this section, we briefly describe the four different voice transformation approaches we explored for speaker de-identification.

A. Voice Transformation based on GMM-mapping (Baseline)

For our baseline system, we used our freely available GMM-mapping based VT system [4] to convert source speakers to a target synthetic voice called kal-diphone [5].

The VT system has a training phase and a testing, or transformation, phase. Training is based on pairs of parallel utterances with the same text spoken by both the source and target speaker. Training collects speaker means and standard deviations for log F0 and computes mel-scale warped cepstral

coefficients (MCEPs) and their dynamic features. The joint distribution of acoustic features (1^{st}-24^{th} MCEPs and their dynamic features) from the source and target speaker is modelled with a GMM. During transformation, the source speaker's log F0 is z-score mapped to match the target mean and standard deviation. Power (0^{th} MCEP) is taken from the source speaker. A detailed description of the training and transformation procedures can be found in [1].

B. De-Duration Voice Transformation (DurVT)

As VT based on GMM-mapping is a frame-by-frame process, the baseline system produces speech that retains the duration characteristics of the source speakers. This might be a disadvantage against SID systems which could use duration characteristics to identify a speaker. We therefore proposed a strategy called DurVT. It tries to normalize transformed speech durations [1]. During training we linearly scale durations of source speakers to match target utterance durations. We then scale the durations of test utterances based on training set statistics.

C. Double Voice Transformation (DoubleVT)

The DoubleVT approach [1] is motivated by our assumption that the baseline voice transformations did not move far enough away from the source speakers. We therefore compose a double VT by applying two VTs in sequence, i.e. we first transform the source speaker to the target speaker (kal-diphone synthetic voice) via de- DurVT and second we transform the de-durationed transformed speech to the target speaker (kal-diphone synthetic voice) via baseline voice transformation.

D. Transterpolated Voice Transformation (TransVT)

The reason we proposed transterpolated voice transformation is that we think double transformations still did not move far away enough from the source speaker. As our baseline VT systems essentially perform linear mappings from the space of source speaker features to the space of the target speaker features, we explored an extrapolation beyond the target speaker. We refer to this process of inter- or extrapolating between the source speaker and converted features as "transterpolation." In this technique, the transterpolated feature x, is computed from the formula $x=s+f*(v-s)$, where s is the value of the source speaker's feature, v is the value of the converted feature, and f is the factor of inter- or extrapolation [1]. The relationship between the factor f and transterpolation can be described as:

- f=0: source speaker (resynthesized)
- 0<f<1: interpolation between source and baseline VT
- f=1: baseline VT
- f>1: extrapolation beyond target speaker

Large factors in transterpolation may project the transformation into non-speech, but values greater than 1.0 still produce normal sounding speech. In conventional voice transformation the goal is to produce output as close as possible to the target speech, but in de-identification we do not require being close to the target, only far away from the source, so transterpolation can be justified.

Though we typically transterpolate both fundamental frequencies and warped cepstra, we decided to experiment with transterpolating the warped cepstra only as it seemed that transterpolated fundamental frequencies might be more exploitable for identifying the source speakers. Our improved GMM-based SID system proved that making use of fundamental frequency related features improves the identification system performance [6]. Therefore, it suggests our future discretion of improving fundamental frequencies transformation for speaker de-identification. Also, as we wanted to focus solely on the contribution of transterpolation to de-identification in this set of experiments, we did not combine it with duration modification, though that would also be possible.

III. SPEAKER IDENTIFICATION (SID)

Using VT for de-identification in automated processes is only as good as the ability of a VT system to deceive a speaker identification (SID) system. If a SID system is able to identify the source speaker from the transformed speech, the speech has not been successfully de-identified. We used two SID systems to evaluate the success of speaker de-identification via voice transformation techniques, a Gaussian Mixture Model (GMM)-based and a Phonetic system. The two systems capture speaker characteristics at different levels: low-level short-term spectral features by the GMM-based system and high-level super-segmental features beyond spectral representations by the Phonetic system.

A. GMM-based SID System

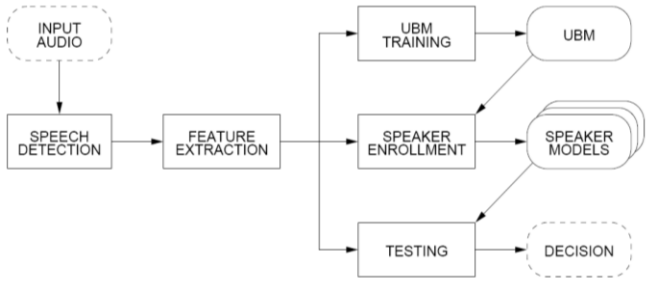

Fig. 1. GMM-based SID System

The GMM-based SID system used in this work is shown in Fig. 1. After speech detection and feature extraction on the input audio it operates in one of three modes: (1) universal background model (UBM) training; (2) speaker enrolment; and (3) testing. All three modes of operation rely on identical feature extraction, which frames the signal every 10ms into 32 ms windows. Frames whose energy is too low to be considered speaker-discriminative are excluded from subsequent processing. From each remaining frame, the first 20 MFCCs are computed and normalized using cepstral mean subtraction (CMS), yielding the final feature vector. Speaker enrolment is preceded by UBM parameter inference [7] via the expectation-maximization (EM) algorithm using a large

corpus of speech. The means of the enrolment speaker model are then adapted based on the UBM via maximum *a posteriori* (MAP) estimation [8], using only the enrolment speaker's speech. Testing proceeds by applying the same feature processing as for model training. The observed sequence of feature vectors is then scored by each speaker's model. The system hypothesizes the speaker whose model best accounts for the observed sequence, *i.e.* gives the highest likelihood.

B. Phonetic SID System

The basic idea of our Phonetic SID system is to apply a statistical model of a speaker's pronunciation, which gets trained on phone sequences that are derived from the speaker's utterances. Although the phone sequences are decoded by phone recognizers using acoustic features, the identification decision is made based solely on the phone sequences [9]. In our Phonetic SID system, phone sequence decoding is performed using Phone Recognizers that are available in 12 languages from GlobalPhone [10]. Phone recognition is performed with a Viterbi search using a fully connected null-grammar network of monophones, thus the hypothesized phone sequence relies on acoustic evidence only.

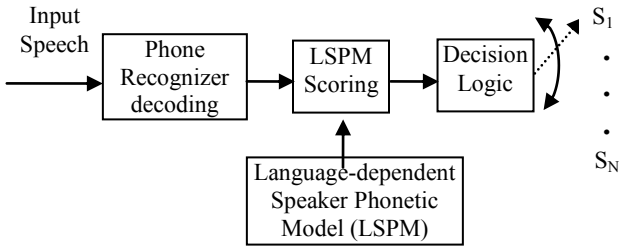

Fig. 2. Phonetic SID System

A Language-dependent Speaker Phonetic Model (LSPM) is generated using the n-gram modeling technique with the CMU-Cambridge Statistical Language Modeling Toolkit (CMU-SLM), *i.e.* for each combination of speaker and phone recognizer a separate bi-gram phone model is trained. Fig. 2 shows the steps of the Phonetic SID system using a single-language phone recognizer:

1. The phone recognizer processes the spoken test utterance to produce a test phone sequence.
2. The perplexity of the resulting test phone sequence is computed based on all previously trained LSPMs.
3. The system hypothesizes the speaker whose LSPM gives the lowest perplexity score on the phone sequence.

This process can be expanded to use multiple phone sequences from a bank of phone recognizers trained on different languages. In our case, each phone stream is independently scored and the scores are fused together with equal weights to form a single decision score. As mentioned above, we apply a bank of 12 parallel GlobalPhone phone recognizers for all experiments in this paper.

IV. Experimental Results

A. Data

The database used for our experiments in this paper consists of data drawn from the LDC CSR-I (WSJ0) [11] and LDC CSR-II (WSJ1) [12] corpora. Speech snippets consist primarily of read sentences from the Wall Street Journal, but also include some spontaneously produced utterances. They were selected from files recorded with a Sennheiser HMD-414 close-talk head-mounted microphone. For each source speaker, VTTrainSET, SIDTrainSET, and TESTSET contributions were constructed by accumulating utterances until there were at least 6 minutes of speech data for training VT systems (VTTTrainSET), 5 minutes of speech data for training SID systems (SIDTrainSET), and at least 3 minutes, in 3 trials of 1 minute each, for testing (TESTSET). In total, this results in a set of 95 male speakers and 102 female speakers. The total number of test trials was 285 and 306, for male and female speakers, respectively. Speech from the remaining speakers (approximately 70 hours) was placed in the UBMSET for UBM training in the GMM-based SID system. 4096 Gaussian mixtures are used for UBM and GMM. We selected the kal-diphone synthetic voice available in the Festival distribution [5] as the target speaker to construct voice transformed versions for each of the male and female speakers.

The baseline performances of the GMM-based and Phonetic SID systems on the original natural speech are shown in Table I.

TABLE I
Speaker Identification Accuracy on the Original Speech with GMM-based and Phonetic SID Systems

	GMM-based SID	Phonetic SID
Male	94%	100%
Female	89%	96%

B. De-identification Performance against SID

In earlier work [1], we evaluated the four voice transformation approaches described in section II on a small pilot database which contains 24 male speakers. On this pilot database, TransVT showed the best de-identification performance. The de-identification performance is measured by *de-identification rate*, which is the percentage of the test trials that are not correctly identified by the automatic SID systems. Also from an earlier human listening test [1], we found that transterpolation with factor 1.6 achieves the best intelligibility. Our first experiment in this paper is therefore to confirm the effectiveness of TransVT approach with factor 1.6 on the large database described in section IV.A. As shown in Table II, TransVT achieved 100% and 95.8% de-identification rates against the GMM-based and Phonetic SID systems respectively for the 95 male speakers. For the 102 female speakers, TransVT achieved 97.7% and 99.0% de-identification rates against the GMM-based and Phonetic SID systems respectively. We also found in our investigation that other choices of factors ranging from 1.2 to 2.0 did not improve over 1.6. Also TransVT outperformed the other three VT approaches described in section II for the purpose of speaker de-identification. Finally, the performance is comparable to the de-identification rates on the small data set

and thus indicates that the TransVT approach with factor 1.6 successfully de-identified the source speakers' identities against automatic SID systems even on a larger dataset that is more typical in size for automatic SID experiments

We noticed that the de-identification performance on the female speakers against the GMM-based and Phonetic SID systems is in the reverse order compared to that on the male speakers. This may be due to the fact that in our experiments we used kal-diphone, a male voice, as the target voice. This may deserve more investigation. It also suggests one avenue for our future work which is to optimize the choice of target voices to improve voice transformation.

TABLE II
TRANSVT WITH FACTOR 1.6 DE-IDENTIFICATION RATE AGAINST TWO SID SYSTEMS

	GMM-based SID	Phonetic SID
Male	100%	97.7%
Female	95.8%	99.0%

C. Human Evaluation

De-identification as a means to securely transmit information without revealing the speaker's identity is only useful if the content of the information is transmitted, *i.e.* the voice is still intelligible for human beings. Consequently, we conducted a human evaluation to investigate the intelligibility. As we found in [1], human listeners are able to correctly identify 100% of the words from the transformed speech produced via TransVT with a factor of 1.6. Our second human evaluation reported here aims to study if the identity of the speaker can be successfully hidden even to listeners who know the speaker very well.

We created three listening tests for the purpose of determining whether humans could properly recognize the de-identified speech. The de-identified speech in these tests was created by TransVT speech from five different source speakers to a single target speaker. The five source speakers were from the awb, bdl, jmk, ksp, and rms sets from the ARCTIC data [13]. These five ARCTIC speakers are all male and represent a range of English accents, including Scottish, Canadian, Indian, and two American varieties. In all cases the target speaker was the kal-diphone synthetic voice, which is a male American voice. As in the other de-identification trials, we used TransVT with a factor of 1.6. The volunteer listeners in this human evaluation were picked from people who personally know the five source speakers.

In the first test the listeners were asked to identify the speakers based on 20 samples of de-identified utterances. For each utterance, the listeners were asked to select one out of the five speaker choices. There were four text-disjunct de-identified utterances for each speaker. The utterances were randomly ordered based on a Fisher-Yates shuffle [14].

In the second test listeners were asked to determine if a pair of utterances was spoken by the same speaker. Each of the 25 utterance pairs consisted of one recording from the ARCTIC database and one de-identified utterance. For each of the five speakers, there were five ARCTIC recordings paired with sample de-identified utterances from each of the five speakers. Again, a Fisher-Yates shuffle was used to randomly order the pairs, and none of the utterances contained the same text.

In the third test listeners were asked to listen to 20 sets of utterances, each consisting of one ARCTIC utterance and 3 de-identified utterances. The listeners had to choose which de-identified utterance was closest to the ARCTIC speaker. For each of the five speakers, four example ARCTIC utterances were used. Different de-identified utterances were used with each ARCTIC example, but one always matched the speaker of the ARCTIC utterance. Again, a Fisher-Yates shuffle was used to randomly order the utterances, and none of the utterances contained the same text. The tests were taken by 5 listeners, all of whom were very familiar with the original speakers, and the databases (in fact we included three of the ARCTIC speakers in the set of listeners). For test 1, listeners correctly identified 26 samples out of 100; chance would be 20%. For test 2, 6 out 25 samples were correctly identified (chance would be 5). For test 3, 36 out of 100 were correct where chance would be 33%. Some of the listeners admitted using non-speech properties to improve their scores, such as background silence properties and silence length.

Given these results we are confident that the proposed TransVT technique successfully de-identifies speakers, even if they are well known to listeners.

D. Distinguishability of De-identified Voices

It is also important to keep the de-identified voices distinguishable from each other. For some applications, it is required to hide the original identity of the speaker in the speech, but at the same time, we still want to be able to discriminate different voices. The de-identification may become easier for larger speaker populations, but preserving the distinguishability becomes harder for larger speaker populations.

To prove the distinguishability of de-identified voices, we ran speaker identification experiments on the de-identified voices. We conducted the transterpolation via TransVT with factor 1.6 on the original training data in SIDTrainSET (5 minutes per speaker for training a speaker model), we then trained speaker models using such de-identified speech. Then closed-set speaker identification experiments were conducted on the de-identified speech using both the GMM-based and Phonetic SID systems. On the small pilot database, 100% and 96% identification accuracy were achieved with the GMM-based and Phonetic SID systems, respectively. On the large database, for the male speakers, 100% and 91% identification accuracy were achieved with the GMM-based and Phonetic SID systems, respectively. For the female speakers, 100% and 96% identification accuracy were achieved with the GMM-based and Phonetic SID systems, respectively. These results show that the de-identified voices are clearly distinguishable.

Table III compares the distinguishability (measured by speaker identification accuracy) between the original voices and the de-identified voices. It is interesting to see that the de-identified voices are easier to distinguish than the original speech for the GMM-based SID system. Apparently, VT

transforms the original speaker's feature space such that the speaker classes become more separable in the transformed space. In contrast to the GMM-based SID system, the distinguishability for the Phonetic SID system remains unchanged for the female speakers but is significantly reduced for male speakers. After comparing the results, we found that 17 out of 95 male speakers cannot be correctly identified on the de-identified voices. Fig. 3 compares the number of correctly identified test trials (3 test trials per speaker) for these 17 speakers based on original vs. de-identified speech. We can see from the figure that only one speaker out of the 17 speakers (speaker 11 in Fig. 3) got totally confused as another speaker. The distinguishability drop of de-identified voices of male speakers for the Phonetic SID systems was caused by a small number of speakers.

The observation in distinguishability change also inspires a new direction of our future work which is to investigate voice transformation approaches such as discriminative approaches that can make the de-identified/transformed voices more distinguishable.

TABLE III
DISTINGUISHABILITY OF THE ORIGINAL SPEECH AND DE-IDENTIFIED SPEECH

	GMM-based SID		Phonetic SID	
	Original	De-identified	Original	De-identified
Male	94%	100%	100%	91%
Female	89%	100%	96%	96%

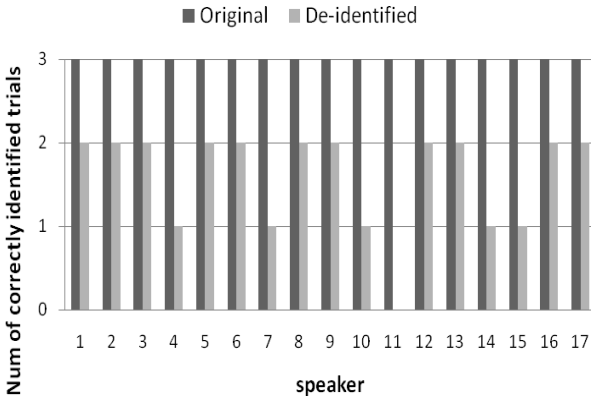

Fig. 3: Number of correctly identified test trials on original vs. de-identified speech by Phonetic SID over the 17 speakers

V. CONCLUSIONS

In this paper we tackle the problem of how to securely transmit information via voice without revealing the identity of the speaker to unauthorized listeners. For this purpose we studied the potential of voice transformation for speaker de-identification. We explored different voice transformation strategies including a standard GMM-mapping based voice transformation, de-duration voice transformation, double voice transformation, and transterpolated voice transformation. The transterpolated voice transformation with a factor of 1.6 gave the best de-identification performance, achieving 100% and 95.8% de-identification rate against the GMM-based and Phonetic SID systems on 95 male speakers and 97.7% and 99.0% de-identification rate against the GMM-based and Phonetic SID systems on 102 female speakers. Human evaluation reveals that factor 1.6 for transterpolation gives full understandability of the securely transmitted content and successfully de-identifies even speech from people who are well known to the listeners.

In the future work, we will investigate improved voice transformation techniques that can successfully de-identify the speakers' identities and at the same time preserve high understandability and naturalness of the speech content and high distinguishability of de-identified voices. We will also explore the impact of these techniques in improving speaker recognition system performance with respect to inherently hard to separate speakers in the original speaker space. As we are interested in preserving the naturalness in the de-identified voices, we would like to also investigate the selection of a wide range of target transformation voices which we feel may make the de-identified voices easier to listen to by humans.

REFERENCES

[1] Q. Jin, A. Toth, T. Schultz, and A. Black, "Voice Convergin: Speaker De-Identification by Voice Transformation", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2009)*, Taipei, Taiwan, 19-24 April, 2009, pp. 3909-3912.

[2] R. Gross, L. Sweeney, F. Torre, and S. Baker, "Model-Based Face De-Identification," *IEEE Workshop on Privacy Research in Vision*, June, 2006, pp. 161 - 168.

[3] Ö. Uzuner, Y. Luo and P. Szolovits, "Evaluating the State-of-the-Art in Automatic De-identification", *Journal of the American Medical Informatics Association*, Vol.14, No.5, 2007, pp. 550-563.

[4] T. Toda, A. W Black, K. Tokuda, "Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, Issue 8, Nov. 2007, pp. 2222-2235.

[5] FestVox: Building Synthetic Voices, 2000. *http://festvox.org*.

[6] K. Laskowski and Q. Jin, "Modeling Instantaneous Intonation for Speaker Identification Using the Fundamental Frequency Variation Spectrum", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2009)*, Taipei, Taiwan, 19-24 April, 2009, pp. 4541-4544.

[7] D. Reynolds, and R. Rose, "Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, Vol.3, No.1, January, 1995, pp.72-83.

[8] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification using adapted Gaussian Mixture Models," *Digital Signal Processing, Vol. 10, No. 1-3, January, 2000, pp. 19-41*.

[9] Q. Jin, T. Schultz, and A. Waibel, "Phonetic Speaker Identification," *Proceedings of the ISCA International Conference on Spoken Language Processing (ICSLP2002)*, Denver, Colorado USA, Sept. 16-20, 2002, pp. 1345-1348.

[10] T. Schultz, "GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University", *Proceedings of the ISCA International Conference on Spoken Language Processing (ICSLP2002)*, Denver, Colorado USA, Sept. 16-20, 2002, pp. 345-348.

[11] J. Garofalo, D. Graff, D. Paul, and D. Pallett, CSR-I (WSJ0) Complete, LDC93S6A, ISBN 1-58563-006-3.

[12] "CSR-II (WSJ1) Complete," *Linguistic Data Consortium,* 1994, vol. LDC94S13A.

[13] J Kominek and A Black, "CMU Arctic Databases for Speech Synthesis", *CMU Technical Report CMU-LTI-03-177* http://festvox.org/cmu_arctic 2003.

[14] R.A. Fisher and F. Yates, "Statistical tables for biological", *in Agricultural and Medical Research* (3rd edition), Oliver & Boyd. London, 1948, pp. 26–27.