VOICE CONVERGIN: SPEAKER DE-IDENTIFICATION BY VOICE TRANSFORMATION

Qin Jin, Arthur R. Toth, Tanja Schultz, Alan W Black
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

ABSTRACT

Speaker identification might be a suitable answer to prevent unauthorized access to personal data. However we also need to provide solutions to secure transmission of spoken information. This challenge divides into two major aspects. First, the secure transmission of the content of the spoken input and second the secure transmission of the identity of the speaker. In this paper we concentrate on the latter, i.e. how to securely transmit information via voice without revealing the identity of the speaker to unauthorized listeners. In order to make the first steps toward solving this problem we study in this paper the potential of voice transformation for speaker de-identification. We use two speaker identification approaches to verify the success of de-identification with voice transformation, a GMM-based and a Phonetic approach, and study different voice transformation strategies to disguise speaker identity information while preserving understandability.

Index Terms— Speaker De-Identification, Voice Transformation, Secure Spoken Information Transmission

1. INTRODUCTION

Speech-driven applications such as telephone-based banking services become an integral part of our everyday lives. As a consequence, there is an increasing demand to prevent unauthorized access to personal data records and to transmit spoken information in a secure way. A technique to address the first problem consists of automatically identifying speakers by their voice and only granting access to information after passing the speaker identification (SID) test. However, this makes voice-based access systems prone to fraud attacks carried out by voice transformation (VT), an automatic technique for mimicking any spoken input in the voice of any target speaker.

In earlier work we had investigated whether state-of-the-art VT technologies can deceive SID systems. In [1] we studied if VT could be used by attackers to transform their voices in order to pretend to be somebody else, and thus retrieve unauthorized access to an SID-secured system. Our results indicated that current VT has the chance to fool a GMM-based SID system but that a Phonetic SID system which uses higher linguistic knowledge from the speaker, such as idiosyncrasies, can effectively discriminate transformed speech from natural speech. It is likely that the future will see a fierce battle between VT and SID systems trying to outsmart each other.

Automatic speaker recognition systems are known to be very sensitive to mismatching training and test conditions [2], and to signal and voice modifications in general. [3] and [4] studied the performance of speaker verification systems on synthesized speech and found it to be significantly harder than the verification on natural speech. [5] showed the harmonic part of the speech signal contains speaker dependent information that can be transformed to mimic another speaker. [6] studied the impact of intentional voice modifications performed by humans and showed that it makes both humans and speaker recognition systems vulnerable. A recent study in [7] investigated the effect of transformed speech on speaker recognition performance and showed that voice transformation can result in drastic increase of the false acceptance rate.

While SID might be an appropriate answer to prevent unauthorized access, we also need a solution to the second problem, i.e. the secure transmission of spoken information. This challenge divides into two major aspects. First, the secure transmission of the content of the spoken input and second, the secure transmission of the identity of the speaker. In this paper we concentrate on the latter, i.e. how to securely transmit information via voice without revealing the identity of the speaker to unauthorized listeners. We propose to use voice conversion to find a transformation of the speaker's voice that meets two requirements: (1) it does not reveal the speaker's true identity to any unauthorized listener (we will refer to this aspect as **speaker de-identification**), and (2) it transmits a key which allows the authorized listeners to back-transform the voice to its original (**speaker re-identification**). Apparently, the solution to this second problem would also be helpful for the task of preventing unauthorized access to personal data. One obvious solution to transmit spoken content without revealing the speaker's identity would be to recognize the spoken words and then apply text-to-speech synthesis. However, this approach has two drawbacks, first it requires full-fledged and error-free speech recognition, and second the transmitted synthesized voice would not allow to recovery the original speaker.

In order to make the first steps toward solving this problem we study in this paper the potential of voice transformation for speaker de-identification, before applying the voice back-transform to re-identify the speaker. De-identification for protecting the privacy of people has been studied in other fields. In [8], the authors studied the de-identification of face images and in [9], the authors use natural language processing approaches to remove personal health information from medical discharge records. The goal of our study is to securely transmit information of "what was said" but to disguise information about "who said it".

2. BASELINE SYSTEMS

2.1 Voice Transformation (VT)

Voice transformation attempts to make speech from a source speaker sound as if it were produced by a target speaker. One strategy for de-identifying speech from various speakers is to transform it so it sounds like it was all produced by the same speaker. For our baseline system, we used a freely available GMM-mapping based VT system [10] to convert 24 male source speakers from the LDC WSJ0 corpus [11] to a target synthetic voice called kal-diphone [10]. Using a synthetic voice is

advantageous during VT training because it removes the need to make additional target speaker recordings for the training set. Basing the following experiments on a freely available VT system and synthetic voice also makes it simpler for others to reproduce our de-identification strategies. The VT system has both a training phase and a testing, or transformation phase. Training is based on pairs of utterances with the same text spoken by both the source and target speakers. In the following experiments 50 training utterances were used for each speaker. Training collects speaker means and standard deviations for log f0. It also aligns source and target speaker frames based on the 0^{th} through 24^{th} warped cepstral coefficients and their dynamic features. GMM parameters are estimated for the joint distribution of the 1^{st} through 24^{th} warped cepstra and their dynamic features over both speakers using the Expectation-Maximization (EM) algorithm. After the GMM is trained, it is used to convert the source speaker speech for realigning the speech and retraining the GMM. A new source speaker utterance is transformed by first estimating its f0 values and 0^{th} through 24^{th} warped cepstra. Then, z-score mapping is used to convert the source speaker log f0 values to the target speaker log f0 values, and a statistical procedure called Maximum Likelihood Parameter Generation (MLPG) with Global Variance (GV) [12] is used to estimate the 1^{st} through 24^{th} warped cepstra for each target speaker frame. The 0^{th} warped cepstra from the source speaker frames along with the estimated target speaker fundamental frequencies and 1^{st} through 24^{th} warped cepstra are used to synthesize a speech waveform using the Mel Log Spectral Approximation (MLSA) filter [13]. Speech transformed in this manner has reasonable quality [12], though it tends to have signal processing artifacts that make it "buzzy." Using VT for de-identification, however, is only good if a VT system can deceive a SID system. If a SID system is able to detect the source speaker, the speech has not been successfully de-identified. Also, as this VT is a frame-by-frame process, the duration characteristics of the source speakers is retained.

2.2 GMM-Based SID System

The Gaussian Mixture Model (GMM) is the most successful statistical model for speaker recognition [14][15]. A speaker's GMM model consists of finite Gaussian distributions parameterized by a priori probabilities, mean vectors, and covariance matrices. The parameters are estimated by the EM algorithm. Our GMM-based SID system consists of five key components: speech detection, feature processing, pattern matching, decision logic, and speaker enrollment. Energy based speech detection aims to remove silence prior to further processing. We extract 13-dimensional Mel-frequency Cepstral Coefficients (MFCC) and apply Cepstral Mean Normalization (CMN) to remove channel effects. The pattern matching component relates MFCC features to stored speaker models and calculates a probability for each model. The identity of the speaker is decided based on the probabilities. However, the system must first be trained to generate speaker models for each speaker, a process commonly referred to as enrollment. In our system, we trained a GMM model with 256 Gaussian mixtures per speaker.

2.3 Phonetic SID System

Significant progress in speaker recognition had recently been made by including high level features such as idiolect, phonetic relations, prosody, and the like [16,17,18]. The basic idea of phonetic speaker identification is to apply a statistical model of a speaker's pronunciation, which gets trained on phonetic sequences that are derived from that speaker's utterances. Although the phonetic sequences are decoded by phone recognizers using acoustic features, the identification decision is made based solely on the phonetic sequences. The rationale of this approach is that phonetic sequences capture a speaker's idiosyncratic pronunciation.

In our Phonetic SID system, phone sequence decoding is performed using Phone Recognizers that are available in 12 languages from GlobalPhone [17]. Phone recognition is performed with a Viterbi search using a fully connected null-grammar network of monophones, thus no prior knowledge is used about any phone statistics. 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). Phonetic speaker identification using a single-language phone recognizer is performed in three steps: Firstly, the phone recognizer processes the spoken test utterance to produce a test phone sequence. Secondly, the perplexity of the resulting test phone sequence is computed based on all previously trained LSPMs. Finally, the speaker identity is decided based on the perplexity scores. 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 to form a single decision score. As described above, we apply a bank of 12 parallel phone recognizers for all experiments in this paper.

3. EXPERIMENTAL SETUP AND RESULTS

3.1 Database Description

For training and evaluation we used audio and transcripts from the WSJ0 corpus available from LDC [11]. We manually processed the transcripts, correcting some errors and removing duplicate sentences. From this processed set, we selected all male speakers who had at least 55 spoken utterances. This resulted in a set of 24 male speakers. The method of speaker de-identification using voice conversion requires data from a target speaker. We selected the kal-diphone synthetic voice available in the Festival distribution [10] as the target speaker to construct voice transformed versions for each of the 24 male WSJ speakers [1]. 50 out of the 55 naturally spoken utterances per speaker were used to train both voice transformation and speaker models. The remaining 5 naturally spoken utterances per speaker were transformed by the trained VT models and used in the de-identification test.

Throughout the description in our paper <ID> denotes the identity of a speaker, S<ID> denotes the model of target speaker <ID> trained with natural speech, while V<ID> refers to the transformed speech of target speaker <ID>. For example: S01 refers to speaker 01 whose model was trained with natural speech. V01 refers to the speech of speaker 01 transformed to the voice of kal-diphone. To study the effects of voice transformation on SID more carefully, we limited ourselves in this paper to closed-set speaker identification experiments, i.e. we use a closed set scenario with 24 male WSJ speakers. We are aware that the number of speakers and the database is small, and thus does not meet the requirements of today's applications. However, our focus is on the investigation of different transformation approaches, on the confusion between natural and transformed speech, and on

providing a proof of concept that these approaches can de-identify the speaker's identity. An extension to larger databases and to open-set identification is planned for the future.

3.2 Standard Voice Transformation

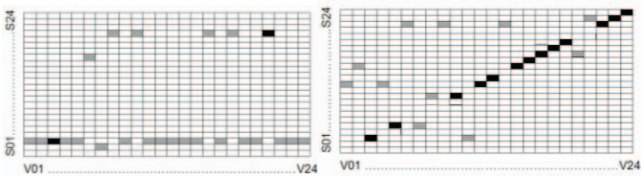

Figure 1: Confusion matrix for GMM-based (left) and Phonetic (right) SID on standard transformed speech, cells in black indicate the trials not de-identified

In our first experiment we transform the voices of all 24 speakers to kal-diphone (V01–V24) and perform speaker identification experiments using the speaker models trained from natural speech of these 24 speakers (S01–S24). In the ideal case of a successful de-identification process, the identification system gives 100% de-identification rate for both identification systems, i.e. the transformed voice cannot be traced back to the corresponding original speaker. Figure 1 shows the resulting confusion matrix between the speaker corresponding to the voice-transformed input speech V<ID> on the x-axis and the hypothesized speaker model S<ID> on the y-axis. The output of the GMM-based SID system is given on the left-hand side, the Phonetic SID system on the right-hand side. The black-colored cells indicate a speaker's match, i.e. despite voice transformation the SID system was able to assign the voice to the original speaker. In other words, black cells indicate when the de-identification was not successful. Figure 1 shows that the GMM system is able to re-identify only two speakers, while the Phonetic SID system correctly re-identifies 13 speakers. The corresponding de-identification rates are 92% and 42% for the standard voice transformed speech with GMM-based and Phonetic SID systems, respectively. Our baseline results show that the standard voice transformation cannot achieve satisfying de-identification performance with the Phonetic SID system. This result is expected based on our previous observations in [1]. So our next focus is to find transformations that are able to prevent re-identification by the Phonetic SID system.

3.3 De-Duration Voice Transformation

As previously mentioned, the baseline voice transformation retains the duration characteristics of the source speakers. Therefore, speaker differences are encoded in the transformation and might be exploited by the SID system to recover source speaker identities. Indeed, there are some questions as to whether these differences in duration may affect the responses of the SID systems. For these reasons, we experimented with a modified version of our baseline transformation strategy with the goal of producing consistent output durations regardless of the source speaker. Training was modified so that the source speaker utterances were scaled to match the durations of the corresponding target speaker utterances using the commonly available sox program which uses a WSOLA algorithm [19]. During training, average duration statistics were calculated for each speaker pair as well. During transformation, we assume the text transcription is not available. (Otherwise, simply synthesizing from the text would be a better de-identification strategy.) As a result, we must rely on the statistics calculated during training to modify the source speakers' utterances before transformation. In particular, we scaled the durations based on the averages. Figure 2 shows the confusion matrix of GMM-based (left-hand side) and Phonetic (right-hand side) SID systems on the de-duration transformed speech. We can see that without durational information the de-identification rates increase, achieving 96% and 46% for the GMM-based and the Phonetic SID systems, respectively.

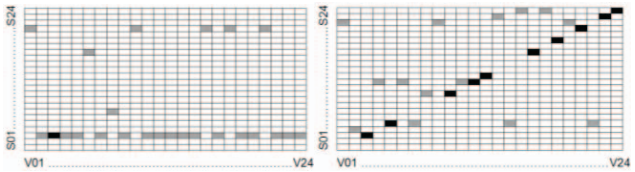

Figure 2: Confusion matrix for GMM-based (left) and Phonetic (right) SID on De-Duration transformed speech, cells in black indicate the trials not de-identified

3.4 Double Voice Transformation

Since there was some measure of de-identification success with the duration-modified VT, we considered the possibility of chaining two voice transformations. The first transformation would be based on the original source and target speakers. The second transformation would then be used to attempt to improve the first transformation by focusing on the differences of the transformed speech with the target speaker. It used output from the first VT for its source speaker and retained the original target speaker. Since the duration-modified voice conversion performed a little better for de-identification than the baseline strategy, it was used for the first VT. As the output of the first VT should then have duration statistics similar to the target speaker, the baseline strategy was used for the second VT. Figure 3 shows the confusion matrix of GMM-based (left-hand) and Phonetic (right-hand) SID systems on the double transformed speech. The de-identification rate for the Phonetic SID system further improved to 67%, while the GMM-based system performance remained unchanged.

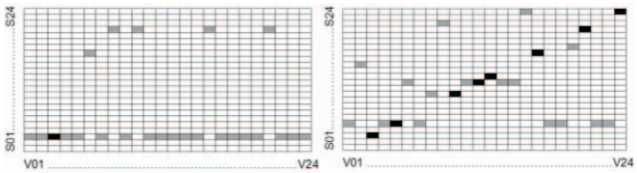

Figure 3: Confusion matrix for GMM-based (left) and Phonetic (right) SID on double transformed speech, cells in black indicate the trials not de-identified

3.5 Transterpolated Voice Transformation

Since the transformed speech in some cases still appeared to have features of the source speaker, we considered the possibility that its identity was in some sense between that of the source and target speaker. As our 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. Though we typically transterpolate both fundamental frequencies and warped cepstra, we decided to experiment with transterpolating only the warped cepstra as it seemed that transterpolated fundamental frequencies might be more exploitable for identifying the source speakers. The following experiments use a single application of transterpolation.

Figure 4 shows the confusion matrix of GMM-based (left-hand) and Phonetic (right-hand) SID systems on the transterpolated speech. Transterpolated voice conversion gave us by far the best de-identification performances, with rates improved to 100% and 87.5% for the GMM-based and the Phonetic SID systems, respectively.

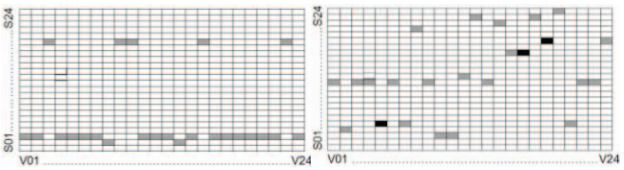

Figure 4: Confusion matrix for GMM-based (left) and Phonetic (right) SID on transterpolated speech with factor 1.6, cells in black indicate the trials not de-identified

3.6 Human Evaluation on Understandability

De-identification as a means to securely transmit information without revealing the speaker's identity is only useful if the content of the transmitted information is still understandable for human beings. Also, in the above experiments we applied transterpolation factors between 1.2 and 2.0. As these factors pushed the cepstra beyond the target speaker's statistics, there was some concern that the quality of the converted speech may become less natural and intelligible. Consequently, we conducted a human evaluation to investigate the understandability and to verify the speaker identity.

Our first test was on speaker identity, where the transterpolated speech was compared to the 3-best ranking potential source speakers. We asked our listeners to identify which speaker the transterpolated speech came from. None of the first three listeners could identify the source speaker (even when they were told the correct answer), thus we refrained from our initial intentions to carry out this experiment with more listeners.

The second test was on intelligibility. A successful de-identification process should preserve the understandability of the transmitted content. We played examples of the de-indentified speech to listeners and asked them to write down what they heard. We then compared this transcript with the reference. We found that it gets harder to understand the transterpolated speech when the factor values increase. For factors 1.2 to 1.6, the listeners can correctly identify 100% of the words, while for factor 2.0, the task is certainly hard, and listeners could only correctly identify about 50% of the words.

4. CONCLUSIONS

In order to make the first steps toward solving the problem of how to securely transmit information via voice without revealing the identity of the speaker to unauthorized listeners, we studied in this paper 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 factor 1.6 gave the best de-identification performance, achieving 100% de-identification rate for the GMM-based and 87.5% for the Phonetic SID system. Human evaluation reveals that factors 1.2 to 1.6 for transterpolation gives full understandability of the securely transmitted content.

REFERENCES

[1] Q. Jin, A. Toth, A. Black, T. Schultz, "Is Voice Transformation a Threat to Speaker Identification?" ICASSP, 2008.

[2] Q. Jin, T. Schultz, and A. Waibel, "Far-field Speaker Recognition," IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No.7, p. 2023-2032, 2007.

[3] B. Pellom and J. Hansen, "An Experimental Study of Speaker Verification Sensitivity to Computer Voice-Altered Imposters," ICASSP, 1999, pp. 837-840.

[4] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using Synthetic Speech against Speaker Verification based on Spectrum and Pitch,", ICSLP, 2000.

[5] D. Genoud and G. Chollet, "Speech Pre-Processing Against Intentional Imposture in Speaker Recognition", ICASLP 1998.

[6] S. Kajarekar, H. Bratt, E. Shriberg, and R. Leon, "A Study of Intentional Voice Modifications for Evading Automatic Speaker Recognition," Odyssey, 2006.

[7] J. Bonastre, D. Matrouf, and C. Fredouille, "Artificial Impostor Voice Transformation Effects on False Acceptance Rates", Interspeech, 2007, pp. 2053-2056.

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

[9] 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.

[10] FestVox: Building Synthetic Voices http://festvox.org 2000.

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

[12] T. Toda, A. W Black, K. Tokuda, "Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory", IEEE TASLP, vol. 15, pp. 2222-2235, Nov. 2007.

[13] S. Imai. 1983. Cepstral analysis synthesis on the mel-frequency scale. In Proceedings of ICASSP 83, pages 93–96.

[14] D. Reynolds, and R. Rose, "Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Trans. on Speech and Audio Processing, vol.3, pp.72-83, 1995.

[15] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification using adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, no. 1-3, pp. 19-41, 2000.

[16] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," Eurospeech, 2001.

[17] Q. Jin, T. Schultz, and A. Waibel, "Phonetic Speaker Identification," ICSLP, 2002, pp. 1345-1348.

[18] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling Prosodic Dynamics for Speaker Recognition," ICASSP, 2003, pp. 788-791.

[19] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech", ICASSP, 1993, pp. 554-557.