IS VOICE TRANSFORMATION A THREAT TO SPEAKER IDENTIFICATION?

Qin Jin, Arthur R. Toth, Alan W Black, Tanja Schultz

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
{qjin, atoth, awb, tanja}@cs.cmu.edu

ABSTRACT

With the development of voice transformation and speech synthesis technologies, speaker identification systems are likely to face attacks from imposters who use voice transformed or synthesized speech to mimic a particular speaker. Therefore, we investigated in this paper how speaker identification systems perform on voice transformed speech. We conducted experiments with two different approaches, the classical GMM-based speaker identification system and the Phonetic speaker identification system. Our experimental results showed that current standard voice transformation techniques are able to fool the GMM-based system but not the Phonetic speaker identification system. These findings imply that future speaker identification systems should include idiosyncratic knowledge in order to successfully distinguish transformed speech from natural speech and thus be armed against imposter attacks.

Index Terms— Speaker Identification, Phonetic Speaker Identification, Voice Transformation

1. INTRODUCTION

Speaker identification (SID) is the procedure of capturing and processing a speech signal and automatically recognizing the speaker who produced the speech. Identifying a person's voice, also called speaker identity in this context, is important for human communication. For example, it allows us to differentiate between speakers in a conference call or on a radio program. Speaker identification technologies have been substantially advanced in the past decades and many recent applications count on reliable automatic speaker identification. However, at the same time, new technologies in the area of automatic voice generation appeared as if they may have the potential to interfere with the advancements in speaker identification. For example, current techniques in speech synthesis can build voices that sound very close to the original speaker, capturing well the style, manner and articulation. Another technique, voice transformation, is designed to modify speech uttered by one speaker such that it sounds like it is spoken by another speaker. Consequently, synthesized or transformed voice can be a serious threat to a speaker identification system or any kind of application that relies on it.

[1] and [2] studied the impact of synthesized speech on speaker verification and observed a significant degradation in system performance. A recent study on intentional voice modifications performed by humans [3] showed that it makes both humans and speaker recognition systems vulnerable. The focus of this paper is to investigate whether current state-of-the-art speaker identification systems can prevent attacks from transformed speech produced by the latest voice transformation technologies.

The rest of this paper is organized as follows. First, we introduce two state-of-the-art SID systems, the classic GMM-based SID system in section 2 and the Phonetic SID system, which uses higher level features, in section 3. Section 4 describes the database, and section 5 presents the experimental setup and results. Finally, we conclude the paper in section 6.

2. GMM-BASED SID SYSTEM

The Gaussian Mixture Model (GMM) is the most prevalent statistical model for speaker recognition [4,5]. A speaker's model based on a GMM consists of a finite number of Gaussian distributions parameterized by their a priori probability, mean vectors, and covariance matrices. The parameters of the model are typically estimated by maximum likelihood estimation, using the Expectation-Maximization (EM) algorithm. A general GMM-based SID system creates a model for each speaker from extracted features in the training phase and then extracts a feature set from input speech of an unknown speaker in the identification phase to decide about the speaker's identity based on all speaker models.

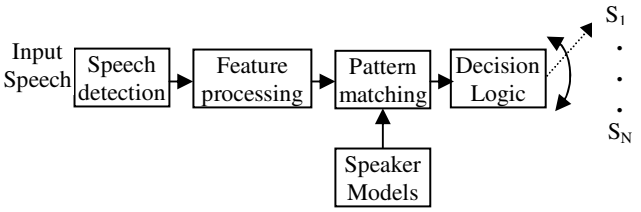

Figure 1. GMM-based SID system components

Our GMM-based SID system as shown in Figure 1 consists of five key components: speech detection (or silence removal), feature processing, pattern matching, decision logic, and enrollment. Speech detection based on the energy of the speech signal is applied to remove silence before further processing. For the feature processing we calculate 13-dimensional Mel-frequency Cepstral Coefficients (MFCC). Cepstral Mean Normalization (CMN) is applied over the MFCC features to remove channel effects. Using these features, the pattern matching component relates them to stored prototypical models and calculates a distortion/probability for each model. The resulting scores are fed into the decision maker, where the system, according to some logic, finally decides on the identity of the speaker. However, the system must first be trained to generate prototypical models for each speaker known to the system, a process commonly referred to as enrollment. In our system, we trained for each speaker a GMM model with 256 Gaussian mixtures.

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 [6,7,8]. 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 utterance. 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 from GlobalPhone in the 12 languages: Arabic (AR), Mandarin Chinese (CH), German (DE), French (FR), Japanese (JA), Korean (KO), Croatian (KR), Portuguese (PO), Russian (RU), Spanish (SP), Swedish (SW), and Turkish (TU). All phone recognizers are trained in the framework of the GlobalPhone project [9]. Phone recognition is performed with a Viterbi search using a fully connected null-grammar network of monophones. Since an equal-probable language model is used in the decoding process, no prior knowledge is used about any phone statistics.

A Language-dependent Speaker Phonetic Model (LSPM) is generated using an n-grams modeling technique. In this paper we estimated bi-gram LSPMs using the CMU-Cambridge Statistical Language Modeling Toolkit (CMU-SLM). As depicted in Figure 2, phonetic speaker identification using a single-language phone recognizer is performed in three steps: Firstly, the phone recognizer processes the test speech utterance to produce a test phone sequence. Secondly, the test phone sequence is compared to all previously trained LSPMs to compute decision scores. Finally, the speaker identity is decided based on the decision scores. This process can be expanded to use multiple phone sequences from a parallel bank of phone recognizers trained on different languages. In this 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. Although the Phonetic system used phone recognizers in multiple languages, it can be applied to any language such as English in this work.

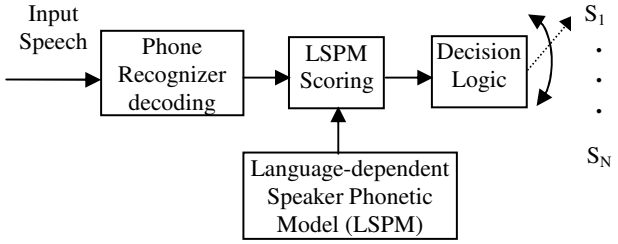

Figure 2. Single language Phonetic SID system components

4. DATABASE DESCRIPTION

For training and evaluating the SID system and as target speakers for voice transformation we used audio and transcripts from the WSJ corpus available from LDC [10]. We intentionally limited our set to male speakers to make the task of speaker identification harder. 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 voice transformation requires data from a source speaker. We selected the kal-diphone synthetic voice available in the Festival distribution [11] as the source speaker to construct voice transformed versions of the 55 utterances for each of the 24 male WSJ speakers. The detailed information of how the voice transformation is produced can be found in [12], in which the speaker identification systems are used for evaluation of the identity of the transformed and synthesized speech. The design of the speaker identification task is such that we have a closed set scenario with the 24 male WSJ speakers. For model training we used 50 utterances per speaker and the remaining 5 utterances for evaluation. The utterance duration ranges from 1 to 20 seconds.

5. EXPERIMENTAL SETUP AND RESULTS

In all our description S<ID> denotes the target speaker <ID> 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 speaker 01 whose model was trained with kal-diphone speech that was transformed to sound like speaker 01.

To carefully study the effects of voice transformation on SID we limited ourselves in this paper to closed set speaker identification experiments. We are fully aware that this is not a realistic scenario for real-life applications but wanted to focus on the confusion between natural and transformed speech first, before taking the next steps of building rejection models for the open-set identification task. For the closed set task we discriminate between two scenarios. In the **single-model** condition, we train a single model per speaker using the natural speech training data. This results in 24 speaker models, which compete when facing natural speech or voice transformed speech from the imposter "kal-diphone". In the **dual-model** condition, we trained two models per speaker, one with the natural speech from that speaker and one with voice transformed speech. This results in 48 competing models and allows us to study if the SID system can discriminate between natural and voice transformed speech.

5.1 Single-model Experimental Results

As a baseline experiment we trained the two SID systems, GMM-based and Phonetic, in the single-model condition with natural speech, i.e. one model per speaker. We tested both systems on 5 sentences natural speech per speaker. Both systems were on par achieving 100% accuracy.

After these baseline experiments, we confronted both systems with the *voice transformed* speech, to simulate an imposter attack. Surprisingly, the two systems show very different behaviors. Figure 3 shows the 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 Phonetic SID system is given at the right in Figure 3, the GMM-based SID system output at the left. The Phonetic SID system hypothesizes mainly two speakers, which as we found in independent experiments later, most closely match the voice of kal-diphone, the *source* speaker of the voice transformation. In contrast, the GMM-based system always hypothesizes the speaker that was used as the *target* speaker of the voice transformation. Thus, our experimental results show that the Phonetic system basically ignores the voice transformation and picks up the identity of the original source, i.e. the imposter speaker, while the GMM-based system picks up the identity of the

target speaker, i.e. the one the imposter is trying to mimic. In other words, the GMM-based system is fooled by voice transformation, while the Phonetic one is not. Based on our current results we can only make this claim for the closed set scenario. In future experiments we will investigate if this also holds in open-set scenarios.

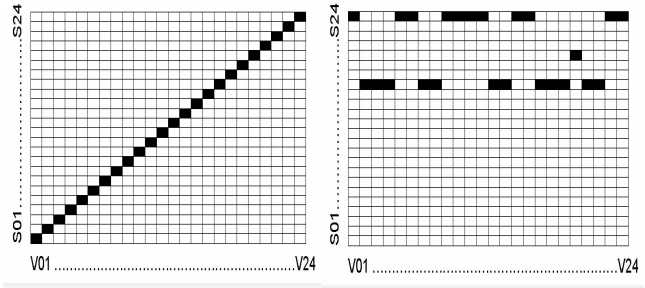

Figure 3: Confusion matrix for GMM-based (left) and Phonetic (right) system with voice-transformed input test speech V<ID> on the x-axis and hypothesized speaker on the y-axis S<ID>

5.2 Dual-model Experimental Results

In the next step we enhanced our speaker identification systems to fight off imposter attacks. This was implemented by training two models for each speaker, one with the natural speech and one with transformed speech. We investigated two test scenarios. In the **natural speech dual-model test** we tested on natural speech and let the 48 dual-models directly compete, thus this task is harder than the 24 single-model task. The purpose of this test is to prove that the SID performance does not degrade compared to the single-model case. In the second test, the **dual-model under attack test** we fed voice transformed speech into the 48-model SID system. This test simulates the situation where the dual-model SID system is facing an imposter attack using voice transformation technology.

For testing the natural speech dual-model condition, we used the same 5 sentences of natural speech per speaker for evaluation as in the single-model case. So the experimental setup is the same as the one for closed set speaker identification with 48 enrolled speakers. Again, both GMM-based and Phonetic SID systems achieved 100% identification accuracy, proving that the additional models do not hurt the overall performance on natural speech.

For testing the dual-model under attack condition, we used 5 sentences of voice transformed speech per speaker for the evaluation. Again, both systems achieved 100% identification accuracy. However, a closer look at the results revealed important differences. In order to highlight these, we examined the top-n hypotheses of both systems. When test speech is natural speech, both systems hypothesize only speaker models, which were trained on natural speech. However, on voice transformed test speech, the two systems have very different outputs. Figure 4 shows the speaker confusion matrix of the top-5 hypotheses of the GMM-based SID system. The rank is indicated by the color intensity and size of the rectangle. The confusion matrix shows that the corresponding target speaker in natural speech always shows up in the top 5 hypotheses (except one case: V04). Figure 5 shows the confusion matrix for the Phonetic SID system. The corresponding target speaker in natural speech never shows up in the top-5 hypotheses. From these results we can conclude that the GMM-based SID system is more likely to be fooled by voice transformed speech than the Phonetic SID system, even under the dual-model strategy.

In all these experiments, we use the same training data for the SID and the voice transformation system. This setup favors an attacking system based on voice transformation. To prove that there is such bias in the current setup, we conducted an experiment in which voice transformation and SID systems used different training sets. Due to data limitation, only 13 of the 24 speakers have enough data to produce alternate training sets consisting of 40 sentences. The experimental results show that both GMM and Phonetic SID systems can prevent attacks from transformed voices slightly better than before. However, the GMM-based SID system still showed the trend to be likely fooled by voice transformation.

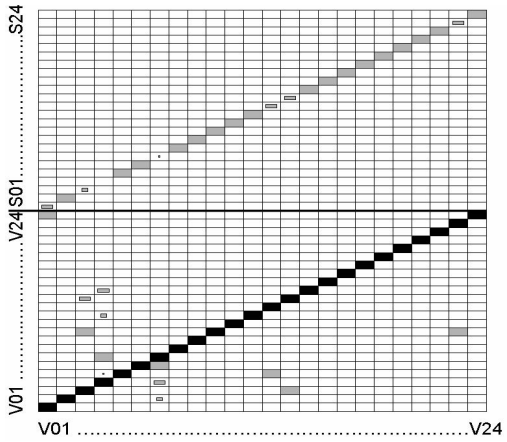

Figure 4: Top 5 confusion matrix for GMM-based SID System in the dual-model under attack scenario (test speaker on x-axis, hypothesized speaker on y-axis; size and grey-level of the rectangle indicate top-N, with full black one = top-1)

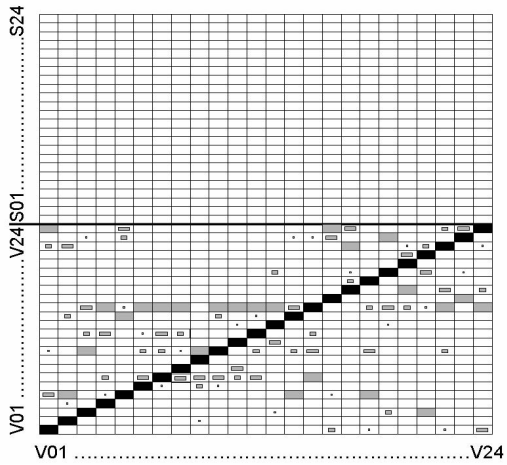

Figure 5: Top 5 confusion matrix for Phonetic SID System in the dual-model under attack scenario (test speaker on x-axis, hypothesized speaker on y-axis; size and grey-level of the rectangle indicate top-N, with full black one = top-1)

5.3 Accuracy versus Sentence Duration

As shown in previous studies [7,13], one of the major limitations of the application of the Phonetic SID approach is that it requires a sufficient amount of training and test data in order to have reliable

performance. We therefore tested the impact of training and test durations on the SID accuracy under the same dual-model condition experimental setup.

Table 1 presents the accuracy of the Phonetic system with a limited amount of training data (from 50 sentences down to 6 sentences). The table shows both the accuracies of the single language system and the combined system which fuses decision scores from all the single language systems together. The results confirm that the amount of training data is a key issue for the success of the Phonetic SID system. We can see from the table that the fusion of multiple languages significantly outperforms each single language especially when training durations get shorter. We also noticed another important fact that when an identification error happens with less training data, the error never happens across the two speech types. It means that when the test speech is transformed speech, if there is an identification error, it always misidentified the test speaker as another target speaker in transformed speech, never as another target speaker in natural speech. It is similar when the test speech is natural speech. This fact proves again that the Phonetic SID system can discriminate transformed speech from natural speech.

Table 1. Phonetic SID accuracy with different training durations

Training Duration	50 Utterances	30 Utterances	20 Utterances	6 Utterances
AR	77.08%	68.75%	75.00%	54.17%
CH	97.92%	89.58%	72.92%	29.17%
DE	85.42%	68.75%	64.58%	52.08%
FR	81.25%	68.75%	68.75%	50.00%
JA	72.92%	62.50%	70.83%	39.58%
KO	77.08%	54.17%	66.67%	62.50%
KR	83.33%	70.83%	54.17%	39.58%
PO	79.17%	75.00%	81.25%	66.67%
RU	81.25%	54.17%	77.08%	52.08%
SP	72.92%	62.50%	70.83%	56.25%
SW	77.08%	89.58%	85.42%	56.25%
TU	87.50%	75.00%	70.83%	56.25%
Combined	**100%**	**97.92%**	**95.83%**	**87.50%**

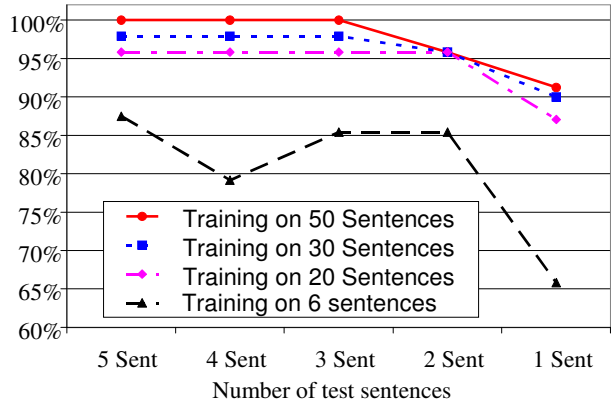

Figure 6. Phonetic SID performance for different test durations

We also tested the impact of test durations on the system performance. Figure 6 summarizes the system performance with different test durations under certain training duration conditions. From the figure we can see that test duration is another key issue for the success of Phonetic SID. We also observed another fact that no cross speech type of error happens when the test duration goes down to 2 sentences. Under the 1 sentence test condition, one or two cross speech type errors happened. The results prove again that the Phonetic SID system can effectively discriminate transformed speech from natural speech.

6. CONCLUSIONS

Is voice transformation technology a threat to speaker recognition? In this paper, we conducted experiments to test whether transformed speech based on current standard voice transformation technologies can attack current start-of-the-art speaker identification systems. We compared two systems, a classic GMM-based SID system and a Phonetic SID system relying on high-level features. Our experiments show that current standard voice transformation has a better chance of fooling a GMM-based SID system than a Phonetic SID system. The phonetic SID system can effectively discriminate transformed speech from natural speech. In the next steps, we will design open-set experiments to investigate this issue more extensively. In the future, we will also conduct experiments to challenge the speaker identification systems with synthesized speech.

REFERENCES

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

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

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

[4] 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, Jan. 1995, pp. 72-83.

[5] 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.

[6] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," Eurospeech, 2001.

[7] Q. Jin, J. Navratil, D. Reynolds, J. Campbell, W. Andrews, and J. Abramson. "Combining Cross-Stream and Time Dimensions in Phonetic Speaker Recognition," ICASSP, 2003, pp. 800-803.

[8] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," ICASSP, 2003, pp. 788-791.

[9] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition," *Speech Communication*, 35:31–51, 2001.

[10] John Garofalo, David Graff, Doug Paul, and David Pallett, CSR-I (WSJ0) Complete, LDC93S6A, ISBN 1-58563-006-3.

[11] http://www.festvox.org

[12] A. Toth, Q. Jin, T. Schultz, and A. Black, "Evaluation of Identity of Synthetic and Transformed Voices Using Speaker Identification Systems," Submitted to ICASSP 2008.

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