ON THE POTENTIAL THREAT OF USING LARGE SPEECH CORPORA FOR IMPOSTOR SELECTION IN SPEAKER VERIFICATION

Johan Lindberg and Mats Blomberg

Royal Institute of Technology (KTH), Department of Speech, Music and Hearing
Drottning Kristinas väg 31, SE-100 44 Stockholm, Sweden, EU
{lindberg, mats}@speech.kth.se
http://www.speech.kth.se/

ABSTRACT

In order to evaluate the risk in SV systems one should take into account the possible perpetrator whom he/she is attacking. This paper thus evaluated if a large speech corpus can be used for selecting impostors to use against a speaker verification (SV) system. We tested the possibility by selecting the most similar speakers from a large corpus and then using the recordings in this corpus for impostor attempts against clients in an SV system. Speech samples of the clients uttering words not used in the SV service were used in order to search the large corpus for similar speakers. Recordings of utterances used in the SV service were then collected from these similar speakers and used for impostor attempts. Our conclusion is that this scenario is a threat that needs to be considered by providers of SV systems.

INTRODUCTION

The aim of this study is to evaluate whether or not large publicly available speech corpora can pose a serious threat to existing speaker verification (SV) systems. A dedicated impostor, able to accurately select subjects from a corpus with similar voice characteristics to the person he/she wishes to deceive, could then use the recordings from the corpus in order to successfully break into the SV protected service. SV systems are often evaluated using randomly selected impostor utterances. That procedure adequately models what would happen in a service where the impostor does not know whom he/she is attacking. This does however not show if SV systems are vulnerable to someone with extensive knowledge about the person to attack as well as the system under attack. The scenario we imagine is that a person, trying to break into an SV protected service, has a speech sample of the target speaker. This utterance does not contain words used in the actual SV service, since, if the perpetrator had recordings of the client uttering words used in the SV protected service he would simply feed these recordings directly to the SV system. The sample is used in order to find speakers in a corpus that are similar to the target speaker. If the corpus also contains recordings of words used in the actual verification, these recordings by the most similar speakers can be used for impostor attempts. Since many SV systems work with digits and many public speech corpora contain digits we expect this to be a feasible approach for an impostor with access to good speech resources. This is a continuation of our work presented in [1]. Other possible scenarios when estimating the threat of dedicated impostors have been studied in [2] and [3].

EXPERIMENT

Speaker Verification system

We used the SV system developed in the European project Picasso [4] for our experiment. The speech was parameterized using 12 LPC-derived cepstral coefficients plus energy, with appended delta and acceleration coefficients (totally 39 elements per frame). The LPC-coefficients are derived from an auto correlation-based LPC analysis of order 16. The analysis window size is 25.6 ms, window shift is 10 ms, pre-emphasis factor is 0.97 and a Hamming window is used. The cepstral parameters are liftered with a liftering factor of 16.

The background model used for calculation of log likelihood ratios was a world model [5]. The world model was trained on 30 speakers, with equal amount of female and male speech used. The world model had the same topology as the client models in all experiments performed. The different client models used are described below.

Databases

We used the Swedish Gandalf database [6] in order to enroll clients into our SV system. Gandalf consists of 82 clients and 131 pseudo impostors. The Swedish FDB5000 SpeechDat database [7] was then used for impostor attempts against the Gandalf clients. SpeechDat contains 5000 subjects recorded in one session each.

The Gandalf database was used with the same division into development, evaluation and world-model speakers as in [8]. This experiment did not use the evaluation set of Gandalf. The decision threshold was set to a speaker dependent threshold at the same-sex equal error rate (EER) point of the development set. The threshold was set with 13-26 true attempts and 40 pseudo impostor attempts per speaker. The test utterances throughout the experiment were all with four digit sequences.

The 40 clients of the Gandalf development set where enrolled using Left-Right Hidden Markov Models (LR-HMM) with one model per digit. Each model had two states per phoneme and 2 gaussians per state. Each client was enrolled using one enrollment session with a total of 25 digit sequences with 5 digits per sequence corresponding to approximately one minute of speech. The enrollment session contained between 10-13 occurrences of each digit. Both enrollment and verification sessions used text-prompted continuously spoken digit
sequences. The enrollment session comes from a call with the favorite handset of the client [6].

For selecting impostors in SpeechDat we used a sentence that was recorded in both corpora. The sentence was “Waxholm ligger i Stockholms skärgård” – (Waxholm is situated in the Stockholm archipelago). The Gandalf development set clients were also enrolled using this fixed non-digit sentence.

The access to identical utterances for the selection procedure is not likely in a practical situation but we wanted to test a worst case. A real perpetrator would probably need to use a text-independent technique in the search for similar subjects and this would further lower his/her chances of finding the right subjects.

The sentence was modeled in two different ways. The first model consisted of an LR-HMM with two states per phoneme and 2 gaussians per state which sums to a total of 56 states and 112 gaussians in the model. The second model consisted of a Gaussian Mixture Model (GMM) with 128 mixtures. Both these models were trained on 5 repetitions of the sentence uttered in one session of Gandalf.

The sentences were recorded in the same session as the digits used for enrollment. This experimental setup monitors an ideal situation. In any real case one would expect the perpetrator not to be able to record the selection utterance in immediate connection to the enrollment. This would lead to a lower success in the imagined scenario than in our experiment, but we designed the experiment in order to test the worst possible scenario.

On the other hand the sentence does not cover the same set of phonemes as the digits. One improvement in the selection criterion would probably be only to use or to increase the importance of those phonemes that occur in the actual SV service when searching the large corpus for similar speakers.

The SpeechDat recordings of the sentence were then used to find the SpeechDat clients that scored best against the Gandalf fixed sentence models. The digits recorded in SpeechDat by these clients were then used for access attempts against the digit models trained from Gandalf. The digit sequence used for the impostor test was “7 9 4 1”, continuously spoken, which was recorded in both corpora.

**EVALUATION METRIC**

Two conditions must be full-filled in order to successfully deceive an SV system with this method. First, there must exist at least one speaker in the corpus that is similar enough to the target client. We will present this as a ratio of target clients that can be matched above the decision threshold for the most similar speaker in the SpeechDat corpus. The second condition is that the perpetrator must select an accepted speaker in a limited number of attempts. We will present this as the ratio of clients deceived using the N most similar (N-best) impostor speakers. This is defined to be the case if at least one of the speakers exceed the decision threshold. We will also study the dependence of this ratio on N. Conventional false acceptance rate (FA) is inappropriate here, since it measures the average rate in the group which will decrease with increasing N.

**RESULTS**

Table 1 shows that the Gandalf development set a posteriori same-sex EER is close to the result obtained when using all the 5000 SpeechDat subjects for impostor attempts. This implies that the speaker distribution is about the same in SpeechDat and Gandalf.

Table 1. False acceptance for the digit models trained on Gandalf development set and tested either against the pseudo impostors in the development configuration of Gandalf or the entire 5000 speaker set of SpeechDat.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>FALSE ACCEPTANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male-Male</td>
</tr>
<tr>
<td>Gandalf dev set a posteriori EER</td>
<td>6.40</td>
</tr>
<tr>
<td>Entire SpeechDat</td>
<td>4.92</td>
</tr>
</tbody>
</table>

The same result is seen when instead looking at the result per client. Figure 1 shows the false acceptance per client with the speaker dependent same-sex EER thresholds set on the Gandalf development set. Figure 2 shows the same case when the entire population of SpeechDat has been used for impostor tests instead. Both impostor populations spread equally well over the population of Gandalf clients even though the SpeechDat population achieves a slightly higher error rate than the Gandalf development set.

Figure 1. False acceptance rates per client for the Gandalf development set with speaker dependent same-sex equal error rate thresholds.
We noticed that when using the GMM or the LR-HMM for modeling the sentence these two would pick the same 10-best subjects in 50% of the cases. This implies that there are discrepancies in the way that the sentence can be modeled but both ways where almost equally successful in picking subjects that were useful in the impostor attempts.

From Figure 3 and Figure 4 it is evident that the selection based on the GMM differs from the selection based on the LR-HMM. The selection processes find different subjects that are successfully used for impostor attempts but averaged over all clients both selection criteria achieve about the same result.

The \textit{a posteriori} selection of SpeechDat subjects shows the upper bound for a perfect selection algorithm. On average this is successful for all but one of the clients (97.5%). As seen in Figure 5 there are several SpeechDat subjects that will score above the threshold for most of the Gandalf clients. It is probably also very hard to detect and prevent this kind of an attack against SV systems. The successful intrusion can be achieved either with a perfect selection criterion or if the service allows a very large amount of trials without blocking the service.

One argument often used for the benefit of SV system is that impostors probably wouldn’t like to give away their own voice in the use of the service. This in order to avoid being tracked when the intrusion is detected. If one considers the fact that an impostor probably would like to avoid using his/her own voice in the SV system, this method would be a feasible way of increasing the chance for getting into the SV protected service.

Figure 6 shows the ratio of deceived clients as a function of the number of impostors used. If the impostor only has one attempt before the service is blocked, then selecting the best subject and using his/her recording will be successful for 20-30% of the clients in Gandalf. If the impostor can test with the 10 best
subjects of SpeechDat then success will be achieved for about 80% of the clients.

The technique demonstrated here only selects the most similar subject in not more than one fifth of the cases. This technique would still be much better than just randomly picking users from a corpus. Depending on how many rejections the actual service allows before alarming the service provider or blocking the service this would be quite enough for successful intrusion.

Figure 6: Rate of deceived clients as a function of the size of the N-best impostor speaker list. An attempt is considered successful if at least one of the speakers exceeds the decision threshold for the target client. The SpeechDat sentences are modeled by a GMM or a LR-HMM.

CONCLUSIONS

The results show that the tested scenario can be a threat to SV systems. The technique is very hard to detect for the service provider, but with some effort in the design of the SV system most of the risk can be removed. One conclusion from this experiment would be the recommendation to avoid words that are common in the large, publicly available corpora when designing text-dependent SV systems.

Furthermore one would have to be careful when using SV systems for people where one easily can collect speech samples from the client. Such persons would be politicians, actors and others who frequently use their voice in media.

Another recommendation would be to combine the security of private passwords and SV systems.

We also recommend that any service protected by SV should also be closely monitored in order to detect strange behavior, such as multiple rejections suddenly occurring for a certain client, even if they occur during separate sessions.

ACKNOWLEDGEMENT

This work was performed within the EU-Telematics program funded project Picasso (LE-8369).

REFERENCES


