VULNERABILITY IN SPEAKER VERIFICATION – A STUDY OF TECHNICAL IMPOSTOR TECHNIQUES

Lindberg J., Blomberg M.
KTH, Dept. of Speech, Music and Hearing
Drottning Kristinas väg 31, SE-100 44 Stockholm, Sweden
{lindberg, mats}@speech.kth.se
http://www.speech.kth.se/

ABSTRACT

This paper reports on some ways to try to deceive a state-of-the-art speaker verification (SV) system. In order to evaluate the risk in SV systems one has to take into account the possible intentional impostors who know whom they are attacking. We defined a worst case scenario where the impostor has extensive knowledge about the person to deceive as well as the system to attack. In this framework we tested our SV system against concatenated client speech, re-synthesis of the client speech and diphone synthesis of the client.

1. INTRODUCTION

In most speaker verification (SV) experiments reported in the scientific literature impostors are chosen randomly from other speakers in a database. This testing method is motivated, from the application point of view, if one assumes that the impostor will not know whom he/she is trying to deceive. With the increasing possibilities to transform speech from a target speaker the question arises on how robust today's SV systems are against this kind of intentional imposture where the impostor can access knowledge about the SV system as well as speech from the client to attack. In order to estimate the security to expect we wanted to try to test a worst case scenario where the impostor has access to target speech and tries to cause an acceptance from the system. This is done in order to estimate an upper bound for how dangerous an impostor can be.

2. SYSTEM AND DATABASE

A state-of-the art Hidden Markov Model (HMM) based SV-system [1] is used as a reference in order to study the imposture techniques. Speech is parameterized with LPCC coefficients and a telephone quality speaker verification database [2] is used in the experiments. This database is split so that the intentional impostor is assumed to have access to one part of it while the other part is used for enrollment to the SV system, thus the impostor is assumed not to have access to the training material used by the SV system. The SV system uses whole word left-right-HMMs with 2 states per phoneme and 2 mixtures per state. The system uses 4-digit sequences as passwords and the sequences used are listed in Table 1. Utterances one to four were repeated several times during the recording of the database, while utterances five to nineteen only occur once or twice per client.

Throughout the experiments one male and one female speaker from the SV database is studied. They both come from the same dialect region (Stockholm) and age group (46-50 years old).

Utterance number	Spoken sequence	Utterance number	Spoken sequence
1	"7 9 4 1"	10	"3 6 1 4"
2	"2 2 3 9"	11	"1 4 9 7"
3	"7 6 8 9"	12	"3 5 9 2"
4	"0 3 5 1"	13	"3 2 1 9"
5	"2 9 5 4"	14	"2 6 5 7"
6	"5 8 3 0"	15	"8 6 1 5"
7	"8 3 5 6"	16	“9 3 2 2”
8	“1 5 3 0”	17	"0 5 4 8"
9	“7 9 4 6”	18	"9 8 6 7"
		19	"0 9 6 7"

Table 1. *The sequences spoken throughout the experiments and their corresponding utterance number as used in the figures. Note that the same sequence is said several times throughout the recordings.*

3. EXPERIMENT

The following impostor experiments were performed:

1 Concatenation of isolated digits being played back to the system

2 Formant copy synthesis of utterances (two different approaches were tested)

3 Telia MBROLA diphone synthesis used to form the requested utterances

The experimental setup neglects any effects due to transmission of the speech signal and concentrates only on to what extent the impostor’s transformation of the voice causes effects that the SV system can distinguish from the client's real speech.

3.1 Experiment one, concatenation of recorded digits

This experiment was set up under the assumption that it would be easy for an impostor to record the genuine client when uttering digits zero to nine and then just cut and paste these recordings to form the correct sequence during verification. The 19 sequences from Table 1 were created from isolated digits uttered by the clients. This speech was then fed to the SV system in order to cause an acceptance.

3.2 Experiment two, re-synthesis

One plausible way for an impostor to attack would be to try and re-synthesise his/her own voice into that of the client through some kind of transform. To show a worst case scenario we tested to re-synthesise the client's own voice in two different ways. This is done under the hypothesis that anything done with the client's own voice must come closer to the model than something done with a different voice. All sequences not used for training the models were re-synthesized and used as impostor utterances.

3.2.1 Re-synthesis 1

The first re-synthesis technique uses a spectral analysis, followed by pitch tracking and an analysis-by-synthesis procedure for estimating the synthesis parameters.

Spectral analysis

Two types of signal analysis are performed. The spectral analysis is performed using a 16-channel Mel filter bank covering frequencies 200 - 6000 Hz. Due to the limited bandwidth of the telephone signal, the two upper channels will have no energy. For pitch tracking a cepstral analysis is also performed. The frame rate is 10 ms in both cases.

Pitch tracking

At each speech frame, a number of pitch candidates are selected from cepstral peaks in the range corresponding to F0 values 50 - 400 Hz. The optimum peak is selected by imposing continuity constraints on the pitch track and its time derivative. This is performed by dynamic programming.

Synthesiser architecture

The cascade formant synthesiser consists of three parallell branches for vocalic, nasal and unvoiced sounds. The LF glottal source model is used as excitation for voiced sounds.

Synthesis parameter tracking

The synthesiser parameters are estimated by a frame-wise analysis-by-synthesis procedure, optionally followed by a piece-wise linear approximation of the trajectories. One purpose of the latter is to reduce the effect of spurious tracking errors.

3.2.2 Re-synthesis 2

The second re-synthesis technique uses a different approach. A text-to-speech system generates rule based parameter value tracks given the label file with manually set segment boundaries. The fundamental frequency, energy and formant frequencies are then extracted from the original speech data. The formant frequency extraction is performed in a two-step procedure. The first formant frequency estimates are obtained from a linear combination of filterbank coefficients [4]. These estimates are fine-tuned by matching a synthetic spectrum with the speech spectrum on a frame by frame basis. This is performed iteratively by changing formant frequencies and energy in small steps in the search for the best match.

3.3 Experiment three, diphone synthesis

In this part of the experiment we used a commercial synthesis system in order to create the requested utterances. The system used was the infovox 330 from Telia Promotor, also known as Annmari. It is trained on studio speech from the client. We tested two scenarios, one where we concatenate isolated digits as produced by the synthesis and one where we let the synthesis produce the exact sequences requested.

4. RESULTS

From figure one it becomes quite clear that the diphone synthesis can be distinguished from the real client for each utterance. The synthesis of the requested sentences perform slightly better than the concatenation of synthesized digits. As seen in Table 2, all of the diphone synthesised utterances can be ruled out without causing a tremendous raise in false rejection of the client.

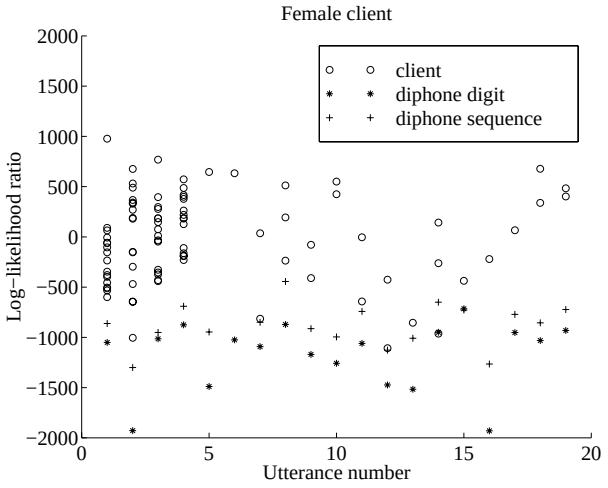

Figure 1. *Concatenation of diphone synthesis digits (star) and diphone synthesis of the uttered sequences (plus) used as impostor utterances against the female client compared with the client's own utterances (circles).*

Impostor type / client sex	EER	FA with EER threshold	FR with EER threshold	FR at FA= 0
random impostor / Female	5.6	5.6	5.6	47.8
random impostor / Male	1.1	1.1	1.1	20.5
diphone digits / Female	5.3	5.3	5.6	5.6
diphone sequences / Female	5.6	38.9	5.6	15.6
concatenated digits / Female	70.0	100	5.6	98.9
concatenated digits / Male	27.3	89.5	1.1	42.6
re-synthesis one / Female	5.6	7.8	5.6	8.9
re-synthesis one / Male	3.4	9.7	1.1	6.8
re-synthesis two / Male	0.6	0	1.1	0.6

Table 2. *Error rates in % for the different ways of attacking the system. The second column is equal error rate calculated with the respective attacking method as impostors and the client's own speech for true client attempts. The second and third column shows the FA and FR when using the threshold from the EER obtained with the randomly chosen impostor attempts, while the last column shows the FR caused in order to rule all technical impostor attempts out.*

The most successful technique was clearly the whole word concatenation. The SV system is unable to distinguish the client's own speech from the concatenated digits. The two re-synthesis versions are however handled better by the SV system and with proper thresholds they would be ruled out without causing any large false reject rates.

5. DISCUSSION

As expected, concatenation of whole words are very effective as a means of impostor while the implemented diphone synthesis and re-synthesising of speech were distinguished from the real speech of the client without causing too many false rejects for the real client. A notable effect is that by just choosing other speakers in the database whom are close to the client's voice and use their recordings for impostor attempts will outperform the re-synthesis we've tried.

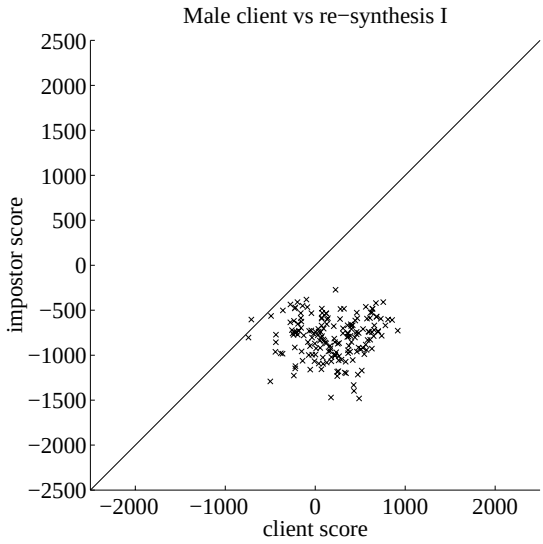

Figure 2. *Client scores for the male client plotted versus the scores of the re-synthesised utterances.*

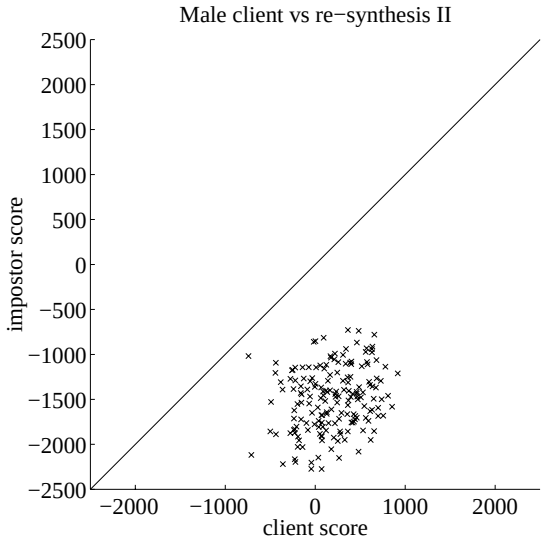

Figure 3. *Client scores for the male client plotted versus the scores of the re-synthesised utterances.*

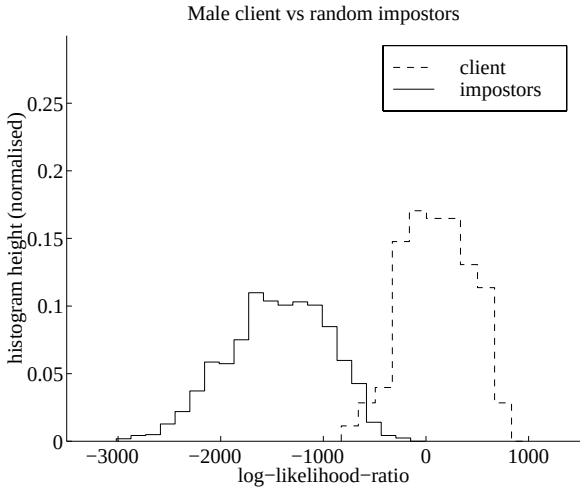

Figure 4. *Histogram of the male client scores and the scores of the randomly chosen impostors*.

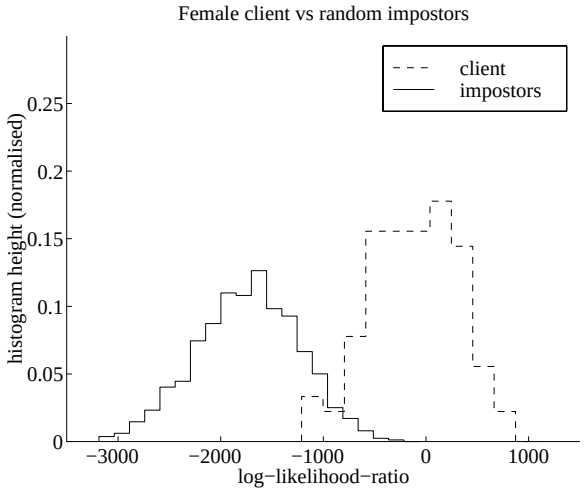

Figure 5. *Histogram of the female client scores and the scores of the randomly chosen impostors.*

There can be several reasons for the fact that the re-synthesis did not come as close to the client model as one would expect. One fact that could give the re-synthesis problems is the coarse spectral resolution. Another disadvantage for the re-synthesis is that it is difficult to track formants when dealing with unknown and varying telephone handsets.

As can be seen in figure 4 and 5 the overlap between random impostor scores and the client score is considerable. This also shows in Table 2 where one can see that an unreasonable amount of false rejects will be created if all the random impostors are to be ruled out. The testing with random impostors thus seems to accurately estimate the security of the system.

Figures 2 and 3 shows the impostor scores versus the client's own score for the two re-synthesis experiments on the male speaker. For re-synthesis one there is a notable effect for a few utterances which give very low scores for the client. For these utterances the re-synthesis actually gives higher scores than the real utterances. One explanation to this could be that these utterances are of very low quality with heavy noise and distortion coming from the telephone handset. These utterances are poorly matched to both the client and the world model and the likelihood ratio obtained is not very accurate.

6. CONCLUSION

One conclusion is that the random testing of speakers against each other actually tells rather accurately what to expect from an SV system even if the impostor tries more intricate techniques. It is however possible with technical impostor attempts and therefore one should evaluate methods for detecting such attempts.

7. ACKNOWLEDGEMENT

This work was done within the EU-Telematics program funded project Picasso (LE-8369). Special thanks to Kjell Gustafson, Telia Promotor for assistance with the diphone synthesis and to Jesper Högberg, KTH, for doing the re-synthesis of the client.

8. REFERENCES

[1] Bimbot F., Hutter H.-P., Jaboulet C., Koolwaaij J., Lindberg J. and Pierrot J.-B. "An Overview of the CAVE Project Research Activities in Speaker Verification". *Proc. Speaker Recognition and its Commercial and Forensic Applications*, Avignon, France, pp 215-220, April 1998.

[2] Melin H. "Gandalf - A Swedish Telephone Speaker Verification Database". *International Conference on Spoken Language Processing*, Philadelphia, USA, pp. 1954-1957, Oct. 1996.

[3] Blomberg, M. "Training production parameters of context-dependent phones for speech recognition", STL-QPSR-1, KTH, 1994

[4] Högberg, J. " Prediction of formant frequencies from linear combinations of filterbank and cepstral coefficients, STL-QPSR-4, KTH, 1997