Automatic Speaker Recognition: Spoofing, Obfuscation and Counter Measures

Hanna Lilja
920112
hanlil@kth.se

Abstract

In many applications of speaker recognition there is a possibility that a user tries to deceive the system. A person might attempt to sound like someone else to either get past an authorisation process or to avoid being recognized. When designing systems which are potential targets of such user behaviour it is essential to take this into account, as to ensure that mistakes by the system are kept to a minimum. There are various ways of fooling a speaker recognition system and also different ways of spotting it. This paper aims to summarise some approaches to both the former and the latter. A recurring theme in the detection methods is the use of the phase spectrum of a speech signal, and the use of temporal information.

1 Introduction

Speaker recognition, as opposed to speech recognition, is the field of research and technology which deals with identifying who is speaking, rather than what is being said. Naturally, this has numerous applications. One such area of application is speaker verification, used as a biometric mean of authentication. Another is speaker identification. In the former case the speaker claims to be a certain person and the task for the system is to determine if the claim is true or not. In the latter, an unknown speaker is to be identified by the system.

With this type of applications one has to take into account the possibility of a user intentionally trying to deceive the system. The nature of the deception depends on what kind of application it is. A speaker verification system is particulary vulnerable to spoofing, meaning that an imposter is trying to prompt a false accept response from the system. There are multiple ways, of varying degrees of sophistication, to achieve this, some more challenging than others to counter when designing the system.

In the case of speaker identification, for example in the context of surveillance, there is a risk that a target takes measures to avoid detection through obfuscation. This means that a speaker disguises or manipulates their speech in order to make the system unable to identify who is speaking, i.e. to provoke a missed detection.

In some aspects, obfuscation is more difficult to deal with than spoofing. One reason is that systems targeted with spoofing often by design has the impostors cooperation in terms of acquiring input such as speech signals, this may not be the case in a surveillance setting and hence the speech signal the system has to work with may be noisy or in other ways of lower quality. Obfuscation is also easier to achieve compared to spoofing, as the former only requires imitation of any voice other than the speaker's own, whereas spoofing requires imitation of a specific speaker.

In this paper recent findings and development related to countering spoofing and obfuscation are discussed. Section 2 briefly describes some models and methods used in speech signal analysis and state-of-the-art speaker recognition systems. Vulnerabilities and counter measures related to spoofing and obfuscation are discussed in section 3 and 4 respectively. This is followed by a general discussion (section 5) and summary (section 6).

2 Background

Automatic speaker verification (ASV) is the task of verifying if a speaker is of a clamied identity or not, i.e. a binary authentication decision where the identity claim is either accepted or rejected. An ASV system can be either text-dependent or text-independent. If the system is text-dependent the speaker must utter a certain pass-phrase in order to be accepted whereas in the case of a text-independet system any phrase can be uttered.

A speech signal can be viewed as quasi-stationary within short time periods (order of magnitude 10^{-2} seconds) and as such a short-time Fourier transform can be applied. The Fourier transform of a speech signal $x(n)$ is on the form:

$$X(w) = |X(w)|e^{j\phi(w)}$$

$|X(w)|$ is refered to as the magnitude spectrum and $\phi(w)$ the phase spectrum. From the magnitude spectrum so called Mel-frequency cepstral coefficients (MFCC) can be obtained. This is done by computing the power spectrum $|X(w)|^2$, applying a Mel-frequency filter bank to the power spectrum and then applying a discrete cosine transform to the logarithm of the output of the filter bank.

When analysing the phase spectrum a useful tool is the group delay function, which measures the nonlinearity of the phase spectrum. It is defined as:

$$\tau(w) = \frac{X_R(w)Y_R(w) + X_I(w)Y_I(w)}{|X(w)|^2}$$

where Y is the short-time Fourier transform of $nx(n)$, and R and I denotes the real and imaginary parts respectively. In order to capture finer details of this spectrum a modified version of the group delay function is sometimes used. A smoothed power spectrum, denoted $|S(w)|^{2\gamma}$, is used instead of the afore mentioned one. The modified group delayed function is defined as:

$$\tau_\gamma(w) = \frac{X_R(w)Y_R(w) + X_I(w)Y_I(w)}{|S(w)|^{2\gamma}}$$

The corresponding modified group dealy phase spectrum is:

$$\tau_{\alpha,\gamma}(w) = \frac{\tau_\gamma(w)}{|\tau_\gamma(w)|}|\tau_\gamma(w)|^\alpha$$

and by applying a discrete cosine transform the phase spectrum cepstral coefficients can be obtained.

One often used method for speaker recognition is a Gaussian Mixture Model (GMM) with a Universal Background Model (UBM). The general idea is to train the UBM with speech samples from a large set of speakers, so that it represents the speaker-independent characteristics of speech. A speaker-specific GMM is trained on speech samples from that particular speaker. For an unkown speaker the likelihood can then be computed with respect to a speaker-specific GMM and the UBM respectively, and the ratio between them serves as score of how well the speakers match. An UBM can also be used as a prior in the training of speaker-specific models.

A common practice in some speaker recognition methods is to put the mean vectors of the speaker specific GMM components together in a GMM supervector. With these supervectors as feature vectors, classification methods such as support vector machines (SVM) with different kernels can be used. As the supervectors may contain information other than speaker related, such as channel information, it is desirable to divide it into components which can each be represented by a low dimension set of factors. This is known as a type of factor analysis (FA). I-vectors are one application of factor analysis used for speaker recognition.

2.1 Other relevant concepts

Vector quantization (VQ) is a method of modelling probability density distributions in data represented as vectors, based on prototype vectors which can be learned from the data or pre-determined. Data is indexed according to the closest prototype vector.

Equal error rate (EER) is a performance metric used for biometric systems (such as an ASV system). It is the rate at which the false match rate (FMR) is equal to the false non-match rate (FNMR). In general the EER can be seen as a threshold corresponding to how similar an input needs to be a template to be considered a match.

3 Spoofing

In this section different spoofing approaches, along with recent findings in terms of counter measures, are discussed. Covered topics are spoofing through artificial signals, converted speech, synthetic speech, and recordings respectively.

3.1 Artificial signals

One relatively simple way of spoofing is the use of artificial signals, that is signals that are tone-like rather than speech-like. The idea is to exploit the fact that certain parts of a speech signal yields higher scores or likelihoods than other parts. A study into the vulnerability of some state-of-the-art ASV systems with respect to this type of spoofing is presented in [1]. In this case the artificial spoofing signals are generated by identifying intervals of speech signals short enough for all frames within them to generate high scores, with the additional requirement that there are enough frames to compute relevant model dependent parameter values. These intervals are then repeated and concatenated to get a signal of sufficient length.

The authors propose and evaluate two approaches to counter spoofing with artificial signals. The first method is based on speech signal features extracted on utterance level. The original model parameters from the speech signal are indexed through vector quantization with respect to the means of the UBM. The index vector can be represented by a histogram, which is reordered and rescaled to produce a new feature vector which describes the whole utterance. For a genuine speech utterance the values of this feature vector will decrease exponentially, whereas for a spoof signal the distribution will look more like a Dirac delta distribution, dominated by the first vector element. The second method is based on voice quality assessment; a state-of-the-art tool designed for this purpose is used. With this tool a mean opinion score ranging from 1 to 5 is computed for each utterance.

The vulnerability to this type of spoofing is evaluated on 3 different ASV systems: a standard GMM system with an UBM, a system based on factor analysis (FA) and an SVM applied to supervectors from the GMM-UBM system. The first two systems are used with 2 different parameter setups, 33 and 50 parameters respectively. The parameterisation with 33 components is used to generate the spoofing signals. The EERs of the different systems under spoofing are compared to their respective baseline EERs, these results are shown in Table 1. As can be seen, ERRs for both the GMM and FA systems are greatly increased under spoofing with only slightly better performance when different parameterisations are used for speaker verification (50) and signal generation (33). The SVM model is inherently robust to this type of spoofing, as shown by the really low ERR under spoofing.

The methods of countering artificial signal spoofing are evaluated by looking at how well they separate spoofing signals from real speech signals. It turns out that voice quality assessment tool is moderately successful in detecting spoofing attempts as the score distribution of spoofing and genuine signals partly overlap, and also causes some false rejection errors. In terms

Model	Parameterisation	Baseline EER [%]	Spoofing EER [%]
GMM	33	8.5	77.1
FA	33	4.8	64.2
GMM	50	7.7	66.3
FA	50	4.2	57.7
SVM	33	7.8	4.1

Table 1: Baseline and spoofing EERs for different systems.

of EER the detection performance is 27%. The utterance level extracted features manage to completely separate spoofing signals from real signals and hence has a 0% EER performance. It is worth noting that the voice quality tool is based only on general speech knowledge whereas the utterance level feature method is less general in terms of what types of artificial spoofing signals it can detect. The authors conclude that while the SVM system and utterance level feature model works well on the particular type of artificial signals investigated, more general countermeasures are needed to protect against unforseen types of spoofing. They suggest the use of frame level score distributions instead of averaged frame scores as a possible way towards making ASV systems more robust to spoofing with artificial signals.

3.2 Converted speech and synthetic speech

An alternativ approach to spoofing is the use of converted speech or synthetic speech. This topic is studied in [2] where an approach based on phase spectrum analysis is proposed, and in [3] where the idea is to distinguish between different types of speech by looking at long-term temporal information derived from both magnitude and phase spectrums. One can exploit the fact that speech synthesis is often done on frame level, where consecutive frames are assumed to be independent of each other, and temporal artifacts may be introduced as a result.

The study presented in [2] is focused on detecting converted speech. Speech conversion is the process of modifying the speech of a source speaker to sound like that of a target speaker. The original speech signal is analysed and parameters like fundamental frequency and spectral envelope parameters are extracted. These parameters are converted according to the target speakers voice and passed on to the synthesis model to create the converted speech signal. In this process the phase information is ignored and therefore lost.

Two different models are used to extract information from the phase spectrum, cosine normalisation and frequency derivative. In cosine normalisation

the phase spectrum is normalised and the discrete cosine transform is applied, coefficient 1 through 12 are kept as features. The frequency derivative is obtained with the help of a modified group delay function. The discrete cosine transform is applied to the output of the group delay function and coefficient 1 through 12 are used. The latter feature set is also refered to as modified group delay cepstral coefficients (MGDCC).

In the experiments performed to evaluate these methods the classification, i.e. the decision if an input signal is natural speech or not, is based on the ratio between log scale likelihoods $log(C \mid \lambda)$ where C is a feature vector and λ is a GMM model for either natural speech or converted speech. Results from three different experimental setups are reported, with different training procedures for $\lambda_{converted}$. With setup A $\lambda_{converted}$ is trained on samples of GMM-based converted speech while with setup B samples of unit-selection based converted speech are used as training data. With setup C it is assumed that no converted speech is available but the analysis and synthesis modules of the conversion model is. The parameters from the speech analysis are passed directly to the synthesis and the resulting reconstructed signal is used as training data for $\lambda_{converted}$.

$\lambda_{natural}$ is the same for all setups. The perfomance of standard MFCC features is reported for comparison. The results in terms of EERs are presented in Table 2. The conclusions are that features extracted from the phase spectrum performs significantly better than standard MFCC in detection of converted speech and that the analyisis-synthesis approach is a viable alternative to converted speech data for training.

Feature model	EER [%] - A	EER [%] - B	EER [%] - C
MFCC	16.80	15.35	20.20
cosine normalisation	6.60	3.93	5.95
frequency derivative	9.13	4.60	2.35

Table 2: Results for detection of converted speech.

In [3] the focus lies on detecting speaker adapted synthetic speech. The proposed model is based on modulation features which capture speech variation across frames. MFCC and MGDCC features, which are both extracted on frame level, are used as a base line in this study. The modulation features can be extracted from both the magnitude and the phase spectrum. The spectrogram is divided into segments of 50 frames, with a shift of 20 frames. A Mel-scale filter bank of size 20 is then applied to each frame in a segment, and the trajectory over frames for each filter bank component is normalized to have mean 0 and variance 1. A modulation spectrum is obtained by applying fast Fourier transform to each of the normalized trajectories and concatenate

each resulting spectra to one vector. Principal component analysis is applied to this vector in order to reduce the dimensionality, the 10 dimensions with largest variance are kept and used as modulation features for the segment. Features obtained from applying the procedure to the magnitude spectrum and phase spectrum are refered to as magnitude modulation (MM) and phase modulation (PM) respectively.

As in the experiments described in [2] the classification is based on the ratio between log scale likelihoods $log(C \mid \lambda)$ where C is a feature vector and λ is a GMM model, in this case for either natural speech or synthetic speech, according to: $\mathcal{L}(C) = log(p(C \mid \lambda_{synthetic})) - log(p(C \mid \lambda_{natural}))$. In order to utilize both short-term spectral and long-term temporal information a combined scoring system is introduced: $\mathcal{L}_{combined}(C) = (1-\alpha)\mathcal{L}_A(C) + \alpha\mathcal{L}_B(C)$ where α is a weighting factor and A and B are two different feature models.

The experiments presented in [3] evaluate the performance of each feature model on its own, as well as the performance of different combinations of features and the influence of α, in terms of EERs. The results for the individual feature models are shown in Table 3. From this it is clear that features derived from the phase spectrum (MGDCC and PM) perform better than the ones derived from the magnitude spectrum (MFCC and MM). It is also noted that the modulation features do not perform well on their own.

Feature model	MFCC	MGDCC	MM	PM
EER [%]	10.98	1.25	19.29	13.71

Table 3: EERs for detection of synthetic speech for individual feature models.

The lowest EER achieved with each combination of two feature models, along with the value of α at which is was obtained, is displyed in Table 4.

Feature models (A + B)	EER [%]	α
MFCC + MGDCC	1.02	0.4
MM + PM	13.33	0.2
MFCC + MM	8.51	0.3
MFCC + PM	7.17	0.5
MGDCC + MM	0.98	0.5
MGDCC + PM	0.89	0.7

Table 4: Lowest EER for detection of synthetic speech for combined feature models.

As can be seen, the best results are obtained by combining spectral features with temporal ones. The temporal features do not contain enough

information on their own, but do hold complementary information such as temporal distortions not captured by the spectral features, which significantly improves the classification.

3.3 Recordings

A third approach to spoofing is the use of recodings, i.e. an impostor playbacks a recording of the target voice to spoof an ASV system. An approach to classify different types of impostors is presented in [4]. The trials a text-dependent ASV system may encounter are divided into four categories: genuine trials (correct pass-phrase, target speaker), naive impostures (wrong pass-phrase, impostor speaker), sly impostures (correct pass-phrase, impostor speaker) and playback imposture (wrong pass-phrase, target speaker). The aim of the study is to classify and counter all types of impostors with the same system and thereby avoid additional computational costs.

A number of verification scores are introduced. The first one is a text-independent score, defined as $S_{ti}(X) = log\frac{L_{\lambda_{gmm}}(X)}{L_{\lambda_{ubm}}(X)}$ where X is a feature vector, L is the likelihood, λ_{gmm} is a speaker-dependent, text-independent GMM and λ_{ubm} a speaker- and text-independent UBM. This can be interpreted as the log likelihood ratio between the hypothesis that the speaker is the target speaker, and the hypothesis that the speaker is an impostor.

Secondly, a text-dependent score is computed, $S_{td} = log\frac{L_{\lambda_{hmm}}(X)}{L_{\lambda_{ubm}}(X)}$, where λ_{hmm} is a speaker- and text-dependent HMM, the likelihood based on the HMM is computed through the Viterbi algorithm.

Thirdly, a speaker normalised score is introduced to specificly target the playback imposture category. This score compares the hypotheis that the target speaker is prounouncing the correct pass-phrase to the hypothesis that the target speaker is pronouncing any pass-phrase, it is defined as $S_{sn} = log\frac{L_{\lambda_{hmm}}(X)}{L_{\lambda_{gmm}}(X)}$.

The imposture classification is done with respect to the S_{td} and S_{sn} scores in a 2 dimensional space. To evaluate the performance of this classification process directly it would be necessary to set fix values for the costs of misclassifications. To avoid doing that a new multi-class score is introduced: $C_{llr} = -\frac{1}{T}\sum_{t=0}^{T} w_t \log_2(P_t)$. P_t is the posterior probability of the true class of a trial t, given a uniform prior, and w_t is a normalising weight factor. This is a positive score measuring the performance of the classifier, the lower the score the better.

The results of two experiments are reported. The first evaluates the individual performance of the three scores S_{ti}, S_{td} and S_{sn} in terms of EERs, the results are displayed in Table 5. In the second experiment the C_{llr} score for

the dual-score classification is compared to C_{llr} for classification based only on the S_{td} score. These results are presented in Table 6. Both experiments are divided into male and female speakers.

	Male			Female		
Impostor type	S_{ti}	S_{td}	S_{sn}	S_{ti}	S_{td}	S_{sn}
Playback	43.48	6.23	0.59	42.99	2.50	0.22
Sly	6.14	1.82	1.90	5.29	0.93	0.88
Naive	5.53	0.59	0.20	4.63	0.12	0.07

Table 5: EERs [%] for detection of different types of impostors.

Classification feature(s)	C_{llr} (Male speakers)	C_{llr} (Female speakers)
S_{td}	0.9429	0.8860
$S_{td} + S_{sn}$	0.6110	0.6325

Table 6: C_{llr} scores for classification based on S_{td} and a combination of S_{td} and S_{sn} respectively.

As can be seen, the S_{sn} score outperformes S_{ti} and S_{td} in terms of EER when it comes to detecting playback impostors, it also improves the detection of naive impostors while maintaining the performance of S_{td} for sly impostors. For the multi-class imposture classification there is lack of equivalent methods to compare with and therefore difficult to do an absolute evaluation. However, there is a significant improvement with the use of dual-score based classification, compared to using only the S_{td} score.

4 Obfuscation

Obfuscation relates to the task of identifying an unknown speaker, for example in a surveillance scenario, and refers to the effort of the speaking target to avoid detection. The vulnerability to obfuscation in speaker recognition system has recieved less attention than the threats of spoofing. Nonetheless it is an important problem and according to [5] obfuscation increases the EER of a standard GMM-UBM system from 9% to 48%, and the EER of a i-vector system from 3% to 20%. The paper presents an assessment of the impact of obfuscation through voice conversion on different ASV systems as well as a new approach to detect obfuscation.

The assessment involve six differnt ASV systems. A standard GMM-UBM system shows the greatest vulnerability to obfuscation, whereas an i-vector system is the most robust one. The EERs with conversion towards the UBM,

a random speaker and the most dissimilar speaker respectively, are shown in Table 7.

Conversion target	GMM-UBM	i-vector
None (baseline)	8.7	3.0
UBM	34.2	8.0
Random	34.5	12.0
Dissimilar	47.7	20.0

Table 7: EERs [%] for different conversion settings.

The proposed approach for obfuscation detection is to exploit the absence of natural spectro-temporal variability in conversed speech through local binary pattern (LBP) analysis of speech spectrograms [6]. This detection method is shown to perform well in the sense that there is almost no degradation of performance for the ASV systems due to obfuscation in the region of low missed-detection rates, i.e. almost all obfuscation attempts are detected.

5 Discussion

Most of the mentioned counter measure methods described are at least to some extent based on an assumption of what type of spoofing or obfuscation a system is targeted with, which affects what anomalies or artifacts in the speech signal may be exploited in order to detect a spoofing or obfuscation attempt. One of the major challenges when applying this type of theory to practice is of course to protect a system against attacks of unpredicted nature. It is also a never ending race against development in other areas of speech technology, in the sense that more sophisticated spoofing and obfuscation methods evolve. For example, as the quality of state-of-the-art speech synthesis improve the more difficult it might become to distinguish it from natural speech.

Keeping up with the technology of potential attackers is one reason why continued research into counter measures is essential, another is the often relatively sensitive nature of the applications of speaker recognition. In a speaker verification system there may be very little room, or indeed no room at all, for mistakes in terms of false accepts. However, a more restrictive system is likely to cause a larger number of false rejects. While the consequences of the latter type of error may be less severe, the occurance of them has to be kept low for the system to be useful. There will always be a trade-off between these two error types, improved spoofing detection makes finding a reasonable balance easier.

6 Summary

There are several ways in which spoofing, i.e. an impostor provoking a false accept response from a speaker verification system, can be performed. Among these are spoofing through artificial signals, converted speech, synthetic speech, and recordings. Different spoofing techniques have different weaknesses which can be exploited when trying to detect a spoofing attempt.

Common tools for this task are features derived from the phase spectrum of a speech signal, and features containing temporal information gathered over time periods significantly longer than typical signal frames. Another recurring concept is that of combining different scores or features that in one way or another cointain complementary information, to improve classification of genuine and impostor trials.

Obfuscation has so far recieved less attention in terms of research, but an approach to detect conversion based on temporal information has been suggested.

References

[1] F. Alegre, R. Vipperla, N. Evans. *Spoofing countermeasures for the protection of automatic speaker recognition from attacks with artificial signals.* Proceedings of INTERSPEECH, ISCA's 13th Annual Conference, 2012.

[2] Z. Wu, E. S. Chng, H. Li. *Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition.* Proceedings of INTERSPEECH, ISCA's 13th Annual Conference, 2012.

[3] Z. Wu, X. Xiao, E. S. Chng, H. Li. *Synthetic Speech Detection using Temporal Modulation Feature.* Proceedings of Acoustics, Speech and Signal Processing (ICASSP) IEEE International Conference, 2013.

[4] A. Larcher, K. A. Lee, B. Ma, H. Li. *Imposture Classification for Text-Dependent Speaker Verification.* Proceedings of Acoustics, Speech and Signal Processing (ICASSP) IEEE International Conference, 2014.

[5] F. Alegre, G. Soldi and N. Evans. *Evasion and Obfuscation in Automatic Speaker Verification.* Proceedings of Acoustics, Speech and Signal Processing (ICASSP) IEEE International Conference, 2014.

[6] F. Alegre, A. Amehraye and N. Evans. *A One-Class Classification Approach to Generalised Speaker Verification Spoofing Countermeasures using Local Binary Patterns.* 2013.