Technical Report RT/28/2013

Towards an Anti-spoofing System Based on Phase and Modulation Features

Maria Joana Correia

INESC-ID/IST

joana.correia@l2f.inesc-id.pt

Advisors:

Isabel Trancoso	Alberto Abad
INESC-ID/IST	INESC-ID

{isabel.trancoso,alberto.abad}@l2f.inesc-id.pt

December 2013

Abstract: Speaker verification systems have seen an increase in popularity in the last years and are being used in broad range of applications. For security reasons it is important to assure the robustness of speaker verification systems against all types of attacks. This report focus on understanding the vulnerability of those systems against a particular type of threat: converted speech. In order to do so, some limitations of state-of-the-art speaker verification and voice conversion systems are addressed and possible counter-measures based on the training of models using phase and modulation features are studied. A baseline speaker verification system is designed and its performance evaluated. The results of the system performance when tested against natural speech and converted speech show equal error rates more than four times bigger for attacks performed by converted speech.

Keywords: Spoofing attack, speaker verification, voice conversion

1 Introduction

Speaker verification (SV) systems have become increasingly popular in the last few years. Their application ranges from simple recreational games to security and biometric applications, which may be protecting confidential information, objects or places from intruders. Some examples of these applications include computer or smartphone log-in, telephone banking or calling cards [1].

With the spread of SV systems in security related applications, the techniques used to try to fool or bypass them have also seen an increase in popularity. One of the possible attacks a SV can undergo is a Spoofing attack, in which an impostor will try to fool the system by converting the characteristics of his voice in such a way that he sounds like the target speaker, thus, boosting his probability of being accepted as the target speaker. Such false positives discredit the use of SV systems in applications that require some level of security.

This report intent is to evaluate the vulnerability of SV systems against spoofing attacks when no anti-spoofing measures are embodied in the system. Then some state of the art anti-spoofing measures that would be possible to include in the baseline system are described. The approach taken is based on the premise that converted speech is different from natural speech of the target speaker, in the sense that the features used to perform voice conversion (VC) describe the speaker in an incomplete fashion. Hence, the converted speech will have inconsistencies and even artifacts when compared to natural speech. The anti-spoofing measures presented are not based on the introduction of new modeling nor detection techniques; rather, they are based on the use of new features that carry information from the phase spectrum of the speech signal and from long term cues of either the magnitude or phase spectrum.

This report is organized as follows: in section 2 and 3 a brief description of the framework of a generic SV system and VC system, respectively, is made. In section 4, the limitations of the features used for SV and VC system as well as some state-of-the-art anti-spoofing measures are presented. In section 5 a description of the baseline SV system is made as well as a comparison with the proposed SV system. The evaluation metrics and the first results are presented in section 6. Conclusions and further work are summarized in section 7.

2 Brief review of Speaker Verification System

The focus of focus this report is on the task of speaker verification. In a verification task the goal is to determine if the identity of a speaker, the *claimant*, is the same as the one the person he claims to be, the *target*. This is a $1{:}\,1$ match task where the claim is simply accepted or rejected.

The typical framework for a generic state-of-the-art SV system comprises two phases: a training phase and a verification phase. In the first one the corpus of speech of the target speaker undergoes front-end processing and feature extraction. Then from those features, a statistical model that characterizes the target speaker is derived. The techniques used to model the speaker identity vary, some examples are Gaussian mixture models (GMM) [2], GMM with universal background model (GMM-UBM) [3], Support Vector Machines using GMM (GMM-SVM) [4], or factor analysis (FA) with its two main variants, joint factor analysis (JFA) [5] and i-vectors [6], among others.

In the verification phase, the claimant utterance undergoes the same type of front-end processing and feature extraction as previously. Then the features are compared to the target speaker model and the decision of accepting or rejecting is performed.

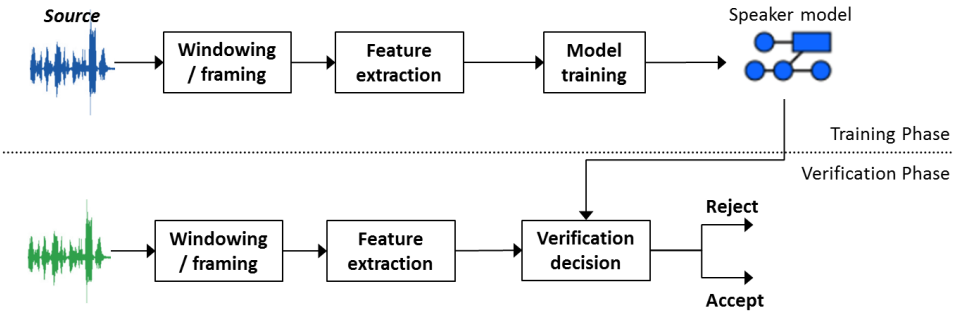

Figure 1. Framework of a generic SV system

3 Brief review of Voice Conversion System

The act of performing VC consists in the adaptation of the characteristics of a source speaker to those of a target speaker in such way that an utterance of the source speaker can be transformed to sound like it was uttered by target speaker, while keeping the original language content [7].

State-of-the-art VC systems can be divided into two categories: the text-dependent and the text independent ones. For the case of text dependent the typical frameworks is somehow analogous to the SV system: The system comprises two phases, the training phase and the transformation phase. In the first one two corpora, one from the source speaker and another from the target speaker are available, they undergo the feature extraction process and a speaker model is estimated from the features. A set of conversion rules between the source and the target speaker models is computed. These conversion rules are used in the training phase to determine the mapping of the features extracted from the test utterance and those of the target speaker. The converted features are then synthesized. A dominant method to perform text-dependent VC is the GMM [8].

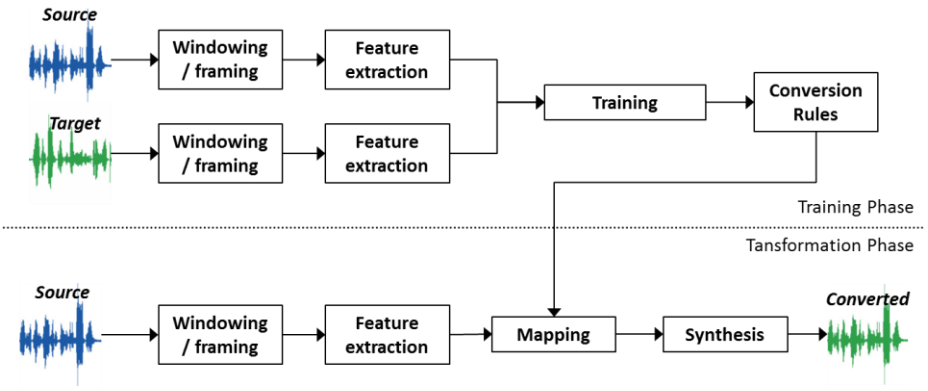

Figure 2. Framework of a generic text-dependent VC system

For text-independent VC, one of the possible approaches unit-section [9]. Systems based in this method use a target speaker speech inventory. To build this inventory the target speaker speech

corpus goes through a feature extraction stage similar to the one of text-independent VC systems. The features extracted from each windowed frame are kept in a database of features of units of speech from the target user. In the transformation stage the features extracted from the test utterance are matched to one of the units of the target speaker by minimizing several cost functions. The units are then concatenated and synthesized.

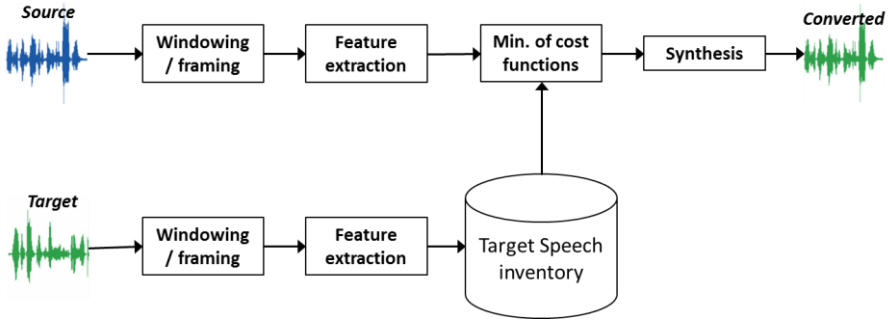

Figure 3. Framework of possible a text-independent VC system

4 Anti-Spoofing Measures

There are many similarities between state-of-the art SV systems and VC systems, two of them are the front-end processing that the utterances used to train speaker modeling techniques undergo and the feature extraction methods as well.

Front end processing in both cases is usually constituted by 3 stages: 1) Pre-emphasis: where a high pass filter is applied to the waveform to emphasize high frequencies and compensate for the human speech production process; 2) Framing: where the utterance is divided in the time-domain waveform into overlapping frames; 3) Windowing: where each frame is multiplied by a window function in order to smooth the effect of using a finite-sized segment for the subsequent feature extraction [10].

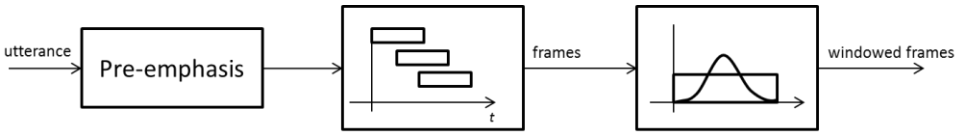

Figure 4. Standard pre-processing of an utterance

The feature extraction process for each windowed frame is also similar for the two aforementioned systems: 1) the Fast Fourier Transform (FFT) of the windowed frame, only the magnitude spectrum is kept, the phase is discarded. 2) Features are derived from the N FFT magnitude coefficients (either MFCCs, RASTA PLP, etc). Usually frames containing silence are discarded [10].

In order to enhance the robustness of the features some sort of normalization and/or channel compensation technique may follow the extraction process.

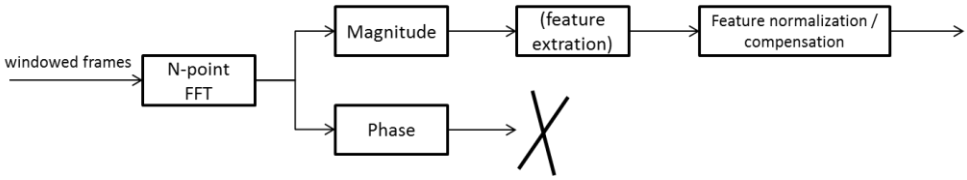

Figure 5. Example of feature extraction process for MFCCs

The features resulting from these two processes are then used to train a speaker model using any of the methods already mentioned.

The choice of the features follows certain requirements for both SV and VC systems. The features should be: 1) a theoretical underpinning so one can understand model behavior and mathematically approach extensions and improvements; 2) generalizable to new data so that the model does not over fit the training data and can match new data; 3) cheap representation in both size and computation [10].

Although the most popular features that were previously mentioned follow, to some extent, the cited requirements, there are limitations to the amount and quality of information that they represent. The most commonly used features, that are derived like described or as a slight variation of it, only comprise information about the magnitude spectrum of the speech signal and, in terms of perception, they only describe the low-level cues of the speech signal, which are related to its acoustic aspects.

The use of these features conveying only short term information of the magnitude spectrum to perform voice conversion leads to an incompletely converted speech signal. The converted speech signal will have inconsistencies and artifacts both in the phase spectrum and in the envelope of the speech signal. By detecting such artifacts one can improve the performance of SV systems against VC attacks.

4.1 Detecting Phase Artifacts

The most common features used in VC systems are derived from the magnitude spectrum The phase spectrum usually is ignored: in a typical VC system the phase spectrum is either not converted at all, or simply discarded and replaced by random spectrum after conversion. While it is true that most of the perceptual information about speech lies in the magnitude spectrum, it has been shown that phase spectrum also carries important information about voice identity [11], thus using information extracted from the phase spectrum to build a speaker models will make them more robust.

From the perspective of a speaker verification system, the performance of the system can be increased by including a module to perform the discrimination based on features extracted from the phase spectrum. In the particular case of a spoofing attack using voice conversion methods the discrimination is simply between natural and converted speech, rather than between the target speaker and any other speaker [12] [13].

4.1.1 The Modified Group Delay (MGD) Phase

To perform the discrimination between natural and converted speech using information from the phase spectrum it is used the modified group delay (MGD) phase spectrum [14] to capture the fine structure of the group delay phase spectrum.

Given a speech signal $x(n)$, the MGD cepstral coefficients can be derived as follows:

i. Compute the short-time Fourier transform (STFT) $X(w)$ and $Y(w)$ of $x(n)$ and $nx(n)$, respectively
ii. Compute the smoothed spectrum $S(w)$ of $|X(w)|$
iii. Compute the MGD phase spectrum $\tau_\gamma(w) = \left(X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)\right)/|S(\omega)|^{2\gamma}$
iv. Reshape the MGD phase spectrum $\tau_{\alpha,\gamma}(\omega) = \left|\tau_\gamma(\omega)\right|^\alpha \tau_\gamma(\omega)/\left|\tau_\gamma(\omega)\right|$
v. Apply the discrete cosine transform (DCT) to the MGD phase spectrum
vi. Keep 12 cepstral coefficients, excluding the 0^{th} coefficient. These modified group delay cepstral coefficients (MGDCC) form a feature vector

Where α and γ are parameters that need to be optimized. After extracting the MGD phase cepstral features, they can be used to train the model of natural vs. converted speech using any of the modeling methods that are successfully used in SV.

4.2 Detecting Long Term Artifacts

Another limitation of the standard features used to perform voice conversion arises from the fact that they are extracted at frame level. These features only comprise information about low level cues and converted speech has very limited information in terms of intonation modulation (among others), which is related to high level cues. Consequently the converted speech will not have the target speaker natural intonation because it has not been converted.

Li proposed the use of modulation of features trajectories in order to discriminate between natural and converted speech [15].

4.2.1 Magnitude and phase modulation

The modulation feature extraction can be accomplished by executing the following steps:

i. Divide the spectrogram (either magnitude or phase) into overlapping segments using approximately a 50 frames window with 20 frames shift (these numbers are rough and their magnitude should be such that provides long term information).
ii. Obtain 20 filter-bank coefficients from the spectrogram, and form a 20x50 matrix.
iii. Apply mean variance normalization (MVN) to the trajectory of each filter bank to normalize the mean and variance to zero and one, respectively.
iv. Apply FFT to the 20 normalized trajectories and compute the modulation spectrum from filter-bank energies
v. Concatenate every modulation spectra on the spectrum and make up a modulation supervector. This supervector can be used as a feature vector

Likewise to the MGDCC, the feature vectors can be used to train a natural vs. converted speaker model.

5 Baseline Speaker Verification System and Proposed Speaker Verification System

Total variability modeling is considered a powerful approach for speaker verification problems while it also has the advantage of representing speech segments in a very compact and efficient way. In this approach, the high dimensional GMM supervector (comprising the speaker and channel variabilities) is modeled as a low-rank total variability space.

The extraction of the low-dimensional total variability factors, or i-vectors, w, from a given speech segment follows:

$$M = m + Tw \tag{1}$$

Where m is the speaker- and channel- independent supervector and T is a rectangular matrix of low rank that represents the primary directions of variability across all training data.

The baseline system used for this report is i-vector based and the matrix m is a GMM-UBM of 1024 mixtures trained using natural speech data. The features used to train the UMB were 49-dimensional vectors of the Energy, 12 MFCCs, 12 deltas and 12 delta-deltas. The adaptation of the UBM was performed by maximum a-posteriori adaptation. The total variability factor matrix, T, is estimated according to [16] and 10 iterations of expectation maximization algorithm are applied.

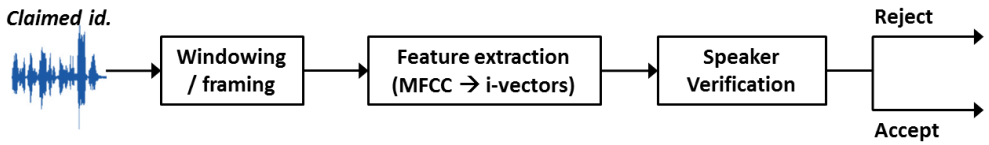

Figure 6. Baseline SV system based on i-vectors

The classification task is performed resorting to a fast scoring method instead of using SVMs without compromising the performance of the system [6]. The score is computed using cosine similarity scores (CSS), which operates by comparing the angles between a test i-vector w_{test} and a target i-vector w_{target}:

$$\text{score}(w_{target}, w_{test}) = \frac{\langle w_{target}, w_{test} \rangle}{\|w_{target}\| \|w_{test}\|} \tag{2}$$

When i-vectors of two speakers point in the same direction they have the highest possible CSS (1) and when i-vectors of two speakers point in opposite directions they have the lowest possible CSS (-1).

The proposed SV system incorporates 2 different converted speech detectors. They are used to test only the utterances accepted by the baseline system, in order to mitigate the high percentage of false acceptances caused by spoofing attacks.

The converted speech detectors are GMM based with 1024 mixtures, for the sake of simplicity. They are trained with converted speech data and the features that are extracted and used to train a natural vs. converted speaker model are the ones described in sections 4.1.1 and 4.2.1, respectively. The decisions are performed using likelihood ratios:

$$\Lambda(X) = \log p(X|\lambda_{natural}) - \log p(X|\lambda_{converted}) \quad (3)$$

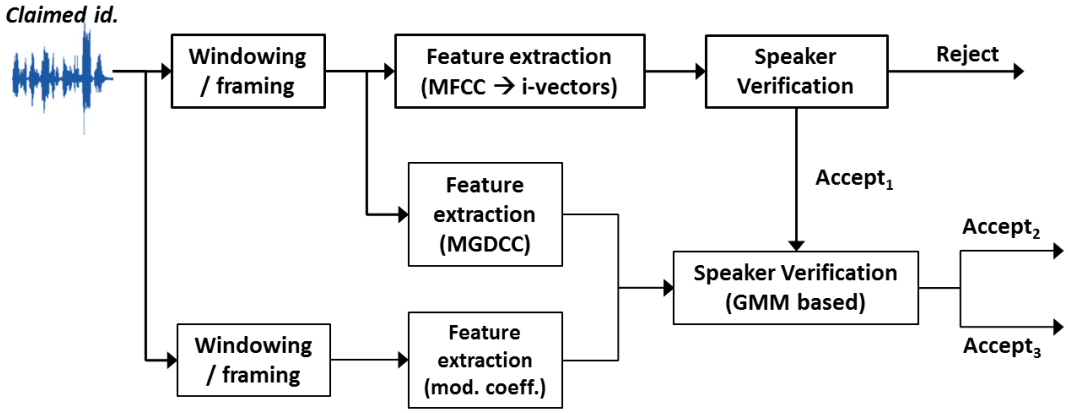

Figure 7. Proposed SV system with anti-spoofing measures

6 Evaluation

The baseline SV system was trained and tested using five different speech corpora, three of natural speech and two of converted speech.

The UBM used to extract the i-vectors was trained using data from subsets of NIST SRE 2004, 2005 and 2006 corpora, restricted to utterances from the enrolment set and verifying the condition 1conv4w-1conv4w where each utterance had an approximate length of 5 min. 1662 utterances of female English speakers and 1520 utterances of male English speakers were selected. Table 1 depicts the number of utterances used broken down by original corpus and sex:

	Female	**Male**
SRE 2004	600	600
SRE 2005	600	600
SRE 2006	462	320

Table 1

The testing of the baseline system was performed using a subset of NIST SRE 2006 corpus, restricted to evaluation set and to the condition 1conv4w-1conv4w where each utterance had an approximate length of 5 min. The subset was used as provided (as natural speech) and was converted using two different VC methods: GMM-based (a text-dependent system) and unit selection (a text independent system). The conversion was performed by Haizhou Li of the Nanyang Technological University, Singapore who kindly provided the converted corpora to us. Table 2 summarizes the test utterances that were used.

	Female	Male
SRE 2006 natural	2083	1564
SRE 2006 GMM	1552	1143
SRE 2006 US	1677	1062

Table 2

The difference between the number of natural and converted utterances is a result of some badly converted files that were not possible to use in the evaluations.

The evaluation metric chosen to measure the performance of the baseline system is the Equal Error Rate (EER). The weights of the cost functions were set accordingly to the NIST 2006 Evaluation Plan. The results for the baseline system are presented in Table 3:

Feature	EER (%)					
	Natural Speech		Converted GMM		Converted US	
	Female	Male	Female	Male	Female	Male
MFCC (baseline)	9.7%	9.6%	46.0%	47.6%	48.5%	49.4%

Table 3

7 Conclusion and Future Work

After designing and implementing a baseline SV system using state-of-the art methods, the evaluation of its performance against natural speech corpora is made. The performance is compared with the performance of the system against two simulated spoofing attacks, using a voice converted by a text-independent VC system and a text-dependent VC system. The predicted effects of spoofing attacks on SV systems were confirmed; the EER rises drastically in both of the spoofing attacks which reduce the applicability of SV systems. The performance of the baseline declines to the point where it is comparable to a random classifier. Such results are consistent with other similar experiments [15].

It is important to mention that this is an ongoing project that is by no means finished. The next step is to implement the anti-spoofing measures described in Section 4 and integrate them in the baseline system like proposed in Section 5. The eventual evaluation of the proposed system will be done using the same metrics and each converted speech detector will be tested against the same 3 corpora summarized in Table 2.

The proposed system is currently being implemented and further evaluations will be made shortly. With the proposed system it is expected a reduction of the EER to levels closer to the EER for the natural speech corpus.

8 References

[1] D. Reynolds, *Automatic Speaker Recognition Using Gaussian Mixture Speaker Models*, The Lincoln Laboratory Journal, 1995

[2] D. Reynolds, R. Rose, *Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models*, IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, January 1995

[3] D. Reynolds, T. Quatiery, R. Dunn, *Speaker Verigication Using Adapted Gaussian Mixture Models*, Digital Signal Processing 10(1-3): 19-41

[4] W. Campbell, J. Campbell, D. Reynolds, E. Singer, P.Torres-Carrasquillo, *Support vector machines for speaker and language recognition*, Computer *Speech and Language*, vol. 20, no. 2-3, pp. 210–229, April 2006

[5] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, *Joint factor analysis versus eigenchannels in speaker recognition*, IEEE Transaction on Audio, Speech and Language Processsing, vol. 15, no. 4,May 2007

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, *Front-End Factor Analysis for Speaker Verification,* IEEE Transaction on Audio, Speech and Language Processsing, vol. 19, no. 4,May 2011

[7] E. Moulines and Y. Sagisaka, *Voice Conversion: State of the Art and Perspectives*, Speech Communication, vol. 16, no. 2, 1995

[8] Y. Stylianou, O. Cappe, E. Moulines, *Statistical methods for voice quality transformation*, Proc. European Conf. Speech Proces. and Techn., pp. 447–450,1995

[9] H. Duxans, D. Erro, J. P´erez, F. Diego, A. Bonafonte, A. Moreno, *Voice conversion of non-aligned data using unit selection,* TC-STAR WSST, 2006

[10] D. Reynolds L. Heck, *Speaker Verification: From Research to Reality,* ICASSP 2001

[11] H. Pobloth, W. Kleijn *On phase perception in speech*, ICASSP 1, 29–32, 1999

[12] P.L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, J. Yamagishi, *Detection of synthetic speech for the problem of imposture,* in ICASSP 2011.

[13] Z. Wu, E. S. Chng, H. Li, *Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition,*Interspeech, 2012

[14] H. Murthy, V. Gadde, *The Modiefied Group Delay and Its Application to Phoneme Recognition,* ICASSP 2003

[15] Z. Wu, X. Xiao, E. Chng, H. Li, *Synthetic Speech Detection Using Temporal Modulation Feature,* ICASSP 2013

[16] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, P. Dumouchel, *A Study of Inter-Speaker Variability in Speaker Verification*, IEEE Transactions on Audio, Speech and Language Processing, vol. 16, no. 5, July 2008