Securing Speaker Verification System Against Replay Attack

Hafiz Malik
Department of Electrical and Computer Engineering University of Michigan-Dearborn, Dearborn, MI, 48128
hafiz@umich.edu

Abstract—In this paper, we present a framework to combat replay attack (RA) on the speaker verification (SV) system. Although the problem of SV system vulnerabilities is not new, however, dramatic improvements in both the SV and the attack models have renewed interest in this area. We model RA using a nonlinear transfer function. Higher-order spectral analysis is considered to capture traces of nonlinearities in the cloned (or replayed) speech of the targeted speaker. We propose scale invariant moments based detection framework to detect cloned audio recording using RA.

Index Terms—Speaker Verification, Security

I. INTRODUCTION

The goal of speaker verification (SV) system is to accept or reject a claim of an identity based on a voice sample. The confirmation of query speech signal subjected to replay attack is a challenging task existing SV systems. This paper presents mathematical modeling of replay attack (RA) nonlinearities. Higher-order spectral analysis is used to capture traces of nonlinearities due to RA and scale invariant moments, e.g., Hu moments [1] are considered to detect cloned speech recording.

A. Problem Statement

Securing SV system against an imposter, e.g., synthetic or altered voice, is one of the most important problems in the SV system design. Many approaches have been proposed [2], [3], [4] to secure SV system against known attacks, e.g., synthetic voice. There has been a little effort to secure SV systems against RA. We are investigating performance of existing SV systems against RA, that is, by cloning a legitimate speaker voice and developing countermeasures to combat such attacks. Shown in Fig. 1 is the conceptual realization of the RA on the SV based access control system.

It can be observed from Fig. 1 that such processing chain for the cloned recording will introduce higher-order nonlinearity (beyond seventh-order) due to cascade of microphone-speaker-microphone (MSM) processing block, assuming 2nd-order nonlinearity for both microphone and speaker.

II. REPLAY ATTACK MODELING

Microphone is an integral component in the processing chain of the RA. Microphones are complex electromechanical devices where interaction between mechanical, electro-mechanical, and electrical elements transforms sound energy into electrical signals. Any nonlinearity in these elements results in a distorted output. In general, the stiffness of the mechanical suspension and acoustical damping are the dominant causes of nonlinear distortion in most microphones. Microphone distortions can be classified into harmonic, and intermodulation distortions. Harmonic distortion is the effect of nonlinearity on a pure tone excitation, consisting of harmonic component in response. The intermodulation distortion is the effect of nonlinearity produced at the output from an excitation consisting of a stronger high frequency \( \omega_2 \) and a weaker low frequency \( \omega_1 \) components. The intermodulation effect produces an output signal made of sums and differences of the input signals fundamental frequencies and their harmonics, that is, \( \omega_2 \pm \omega_1, 2\omega_1, \omega_2 \pm 2\omega_1, \omega_2 \pm 3\omega_1, \) etc.

Microphone response is generally characterized in terms of physical parameters of the microphone [5]. Consider, for example, a condenser microphone with capsule capacitance \( C_o \), input resistance \( R \), diaphragm displacement, \( \delta \), supply voltage \( V_o \), and multisinusoidal excitation signal \( x(t) = \sum_i A_i \sin(\omega_i t) \), where \( \sum_i |A_i| \leq T/4, i \in \{r,s,q\} \). \( w_i C_o R \gg 1 \) and \( 2\delta \omega_i C_o R \ll 1 \). The microphone output \( y(t) \) (by neglecting 4th-order and beyond harmonic and intermodulation terms) can be expressed as,

\[
y(t) = \frac{V_o}{2} \sum_i \delta_i \sin(\omega_i t) + \frac{V_o}{RC_o} \sum_i \frac{\delta_i^2}{\omega_i} \sin(2\omega_i t) \]

\[
+ \frac{V_o}{2(RC_o)^2} \sum_{i,j} \frac{\delta_i^3}{\omega_i^2} \sin(3\omega_i t) \]

\[
+ \frac{\omega_1^2 + \omega_2^2}{\omega_1 \omega_2 RC_o} V_o \sin ((\omega_r \pm \omega_s)t) \]  \( (1) \)

\[
+ \frac{2\omega_r \omega_s}{\omega_1 \omega_2 (RC_o)^2} V_o \sin ((2\omega_r \pm \omega_s)t) \]

\[
+ \frac{2\omega_r \omega_s}{\omega_1 \omega_s (RC_o)^2} V_o \sin ((\omega_r \pm \omega_s \pm \omega_q)t) \]

\[
+ \frac{2\omega_r \omega_s}{\omega_1 \omega_s (RC_o)^2} V_o \sin ((\omega_r \pm \omega_s \pm \omega_q)t) \]

\[
+ \frac{2\omega_r \omega_s}{\omega_1 \omega_s (RC_o)^2} V_o \sin ((\omega_r \pm \omega_s \pm \omega_q)t) \]

\[
+ \frac{2\omega_r \omega_s}{\omega_1 \omega_s (RC_o)^2} V_o \sin ((\omega_r \pm \omega_s \pm \omega_q)t) \]
The microphone response given in Eq. (1), can be approximated using following time-invariant Hammerstein series model,

\[ y(t) = \sum_{\tau} g_1(\tau)x(t - \tau) + \sum_{\tau} g_2(\tau)x(t - \tau)^2 \]  

It can be observed from Eq. (2) that nonlinearity introduced by a microphone can be described with a simple point-wise operation which introduces higher-order correlations. The MSM processing chain, therefore, can be modeled using a higher-order (beyond fourth-order) nonlinear system.

The higher-order spectral analysis is used to capture nonlinearities due to RA. Main motivation behind considering higher-order statistics (also known as cumulants, and their associated Fourier transforms, e.g., polyspectra) is that these methods reveal both the amplitude and the phase information about a process. In addition, cumulant-based methods are blind to any kind of a Gaussian process.

To capture nonlinearities due to RA, we focus on intermodulation distortion-spread, as intermodulation distortion is more dominant in RA output. To verify this claim, we computed bicoherence of the speech recording and the corresponding cloned recording. Shown in the left panel of Fig. 2 is the bicoherence magnitude plot of an audio recording and in the right panel is the bicoherence magnitude plot of the cloned recording.

\[ \eta_i,j = \frac{\mu_i,j}{\mu_{0,0}} \]  

where, the central moment \( \mu_{i,j} \) can be computed as,

\[ \mu_{i,j} = \sum_x \sum_y (x - \bar{x})^i(y - \bar{y})^j b(x, y) \]  

where \( \bar{x} \) and \( \bar{y} \) are the components of the centroid of bicoherence magnitude, \( b(\cdot, \cdot) \).

IV. RESULTS

Effectiveness of the proposed method is tested on a set of audio recordings captured using three commercial grade microphones. To simulate replay attack (RA), each recording was cloned by re-recording it (using same microphone) by playing it through a commercial grade speaker. To investigate impact of \( n^{th} \)-order RA, i.e., the speech signal is subjected \( n \) RAs. For example, \( 2^{nd} \)-order RA is simulated by passing a recording through \( SM - SM \) processing chain. Similarly, \( 3^{rd} \)-order RA is obtained by passing a recording through \( SM - SM - SM \) processing chain. The scale invariant moment \( \eta_{2,0} \) is computed from the resulting cloned recording using Eq. (3). Shown in the Table are the second scale invariant Hu moment \( \eta_{2,0} \) computed from the speech recording and \( 1^{st}, 2^{nd} \) and \( 3^{rd} \) order RA outputs.

<table>
<thead>
<tr>
<th>Mic</th>
<th>Direct</th>
<th>1\textsuperscript{st} order RA</th>
<th>2\textsuperscript{nd} order RA</th>
<th>3\textsuperscript{rd} order RA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>62.5003</td>
<td>114.8635</td>
<td>302.2956</td>
<td>505.0381</td>
</tr>
<tr>
<td>2</td>
<td>85.9233</td>
<td>91.6101</td>
<td>302.5650</td>
<td>476.5635</td>
</tr>
<tr>
<td>3</td>
<td>54.8694</td>
<td>98.2582</td>
<td>227.3322</td>
<td>491.2536</td>
</tr>
</tbody>
</table>

It can be observed that RA does change the \( \eta_{2,0} \), and \( \eta_{2,0} \) increases with increasing order of RA. This characterization of the bicoherence magnitude can be used to distinguish between a legitimate speaker and a cloned recording. We are currently investigating performance of the proposed scheme for existing SV systems against RA. We are also investigating characterizations of higher-order (beyond fourth order) scale and rotation invariant Hu moments for both the magnitude and the phase the bicoherence of speech recording.

REFERENCES