Securing Speaker Verification System Against Replay Attack

Hafiz Malik

Information Systems, Security, and Forensics (ISSF) Lab
Department of Electrical and Computer Engineering
University of Michigan – Dearborn

DEARBORN
COLLEGE OF ENGINEERING & COMPUTER SCIENCE

Outline

- Introduction
- Applications of Cloned Recording
- Replay Attack Distortion Modeling
- Nonlinear Distortion Identification using HOSA
  - Invariant Moments
- Experimental Results
- Conclusion

Speaker Verification System

- 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 (or cloned recording) is a challenging task for existing SV systems.
- This work 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 Hu moments are considered to detect cloned speech recording.

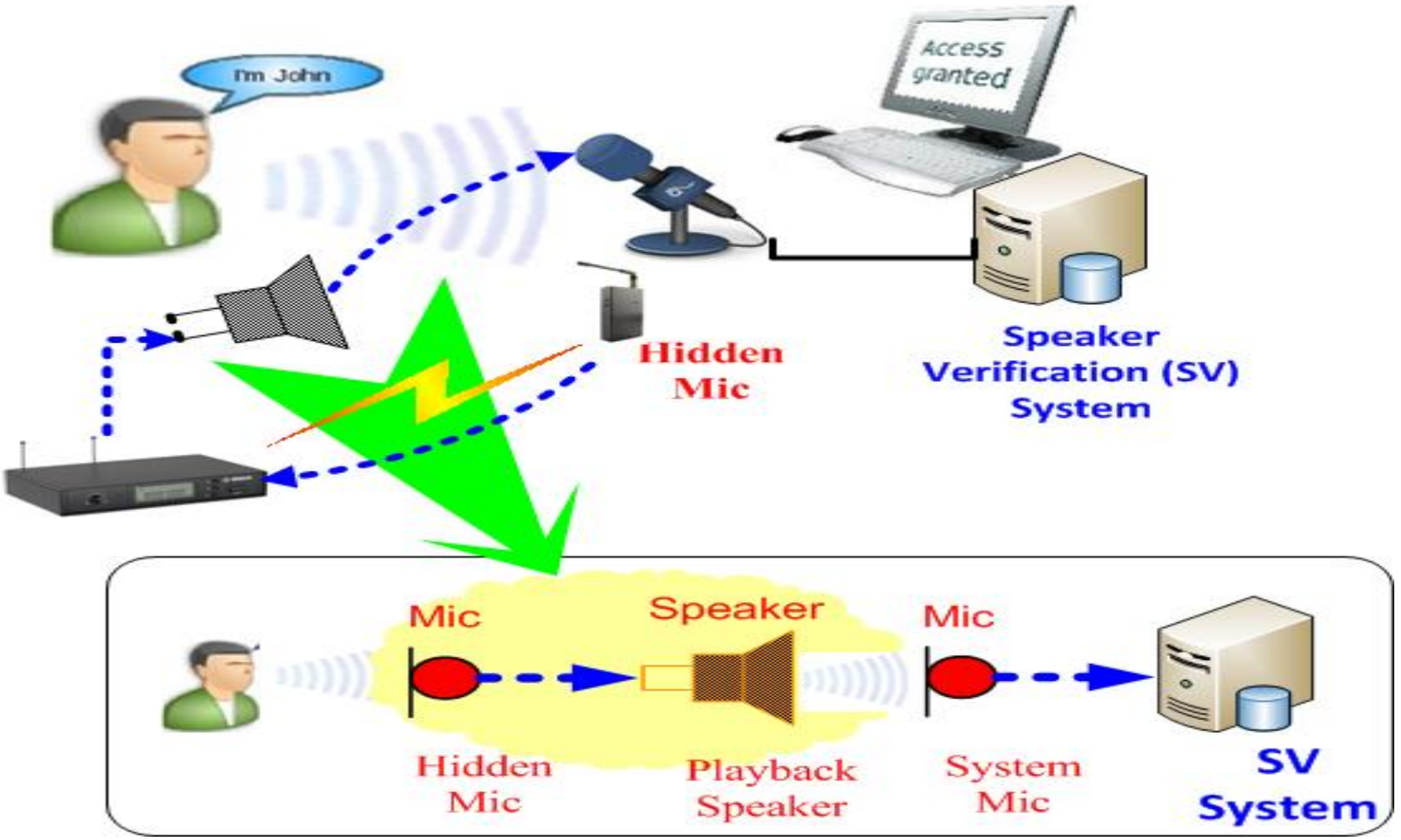
Applications of Cloned Recording Identification

1. Secure Speaker Verification System Design
2. In-Theater Piracy (a.k.a. *camcorder theft*) Detection
3. Robocall Detection
4. Multimedia Forensics, etc.

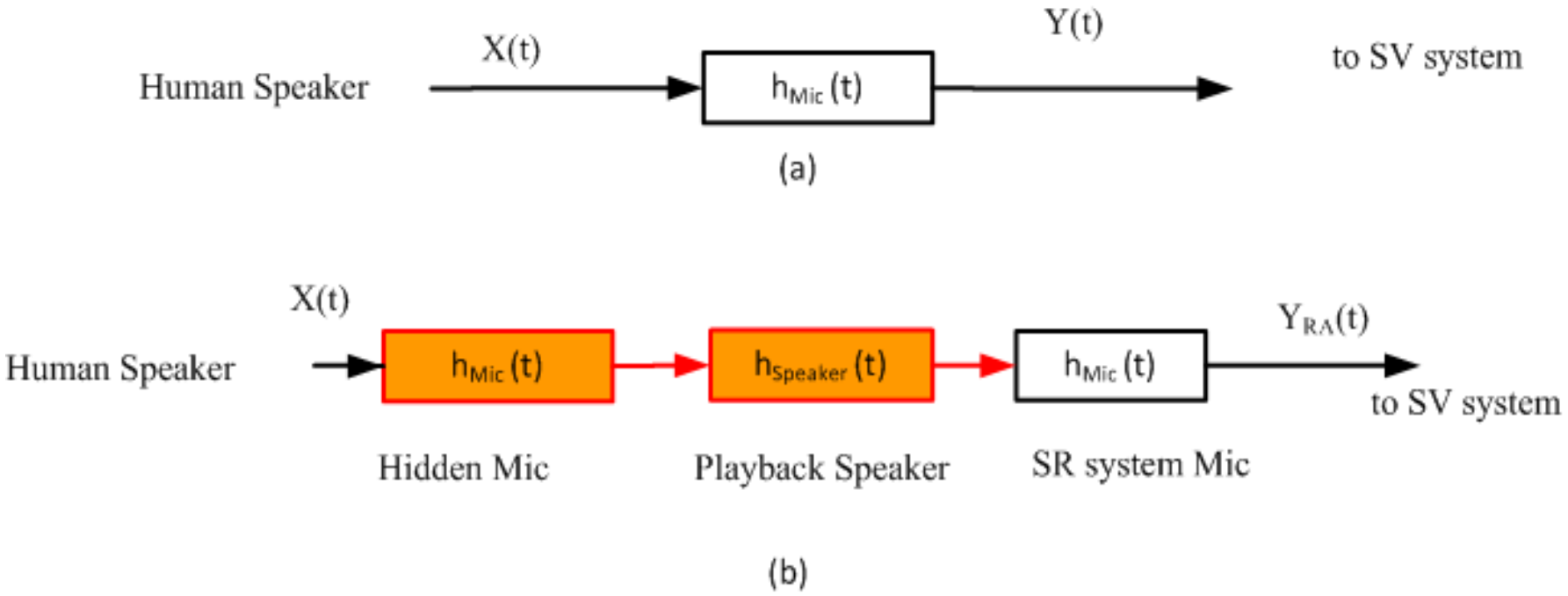
Microphone Forensics: State-of-the-Art

- De-Leon et al. (2010) [synthetic speech]
- Pellom and Hansen (1999) [synthetic speech]
- Masuko et al. (1999) [synthetic speech]
- Rosenberg and Parthasarathy (1996) [background noise]

Conceptual Realization of Replay Attack



SV Processing Chain with & without RA



Microphone Distortion Modeling

- Microphone distortions can be classified into

  1. Harmonic distortion,
  2. Intermodulation distortion, and
  3. Difference-frequency distortion.

- The intermodulation effect produces an output signal made of sums and differences of the input signals fundamental frequencies and their harmonics, that is,

$$\omega_2 \pm \omega_1,\ \omega_2 \pm 2\omega_1,\ \omega_2 \pm 3\omega_1,\ etc.$$

Microphone Distortion Modeling

- The microphone response can be approximated using the following discrete time-invariant Hammerstein series model,

$$y[n] = \sum_k g_1[k]\,x[n-k] + \sum_k g_2[k]\,x^2[n-k] + \sum_k g_3[k]\,x^3[n-k] + \cdots$$

where $g_i[k]$: i=1, 2, 3 characterize linear, quadratic, and cubic response of the microphone

Replay Attack Distortion Modeling

- The nonlinearity introduced by a microphone can be described with a simple point-wise operation which introduces higher-order correlations.

- The RA modeled using Mic-Speaker-Mic (MSM) processing chain, therefore, can be modeled using a higher-order (beyond fourth-order) nonlinear system.

where $g_i[k]$; i=1, 2, 3 characterize linear, quadratic, and cubic

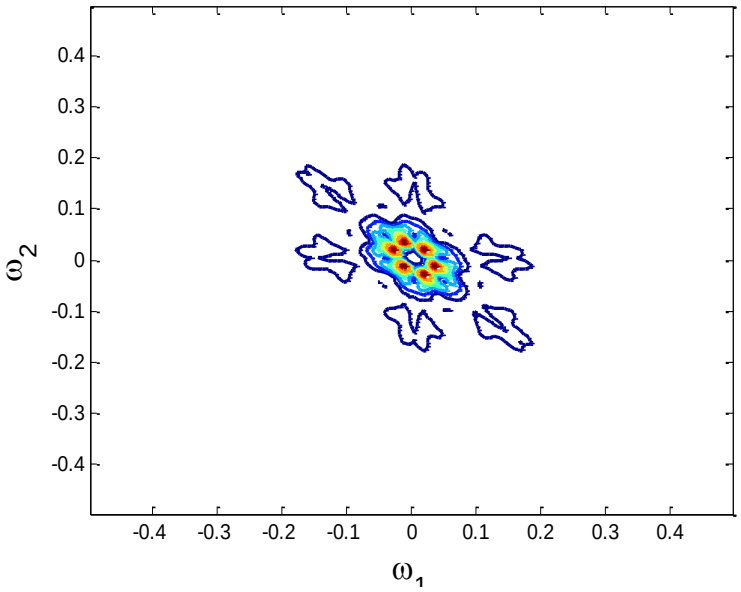
Higher-Order Spectral Analysis to Capture RA Artifacts

- The HOSA can be used to analyze the nonlinearity of a system operating under a random input.

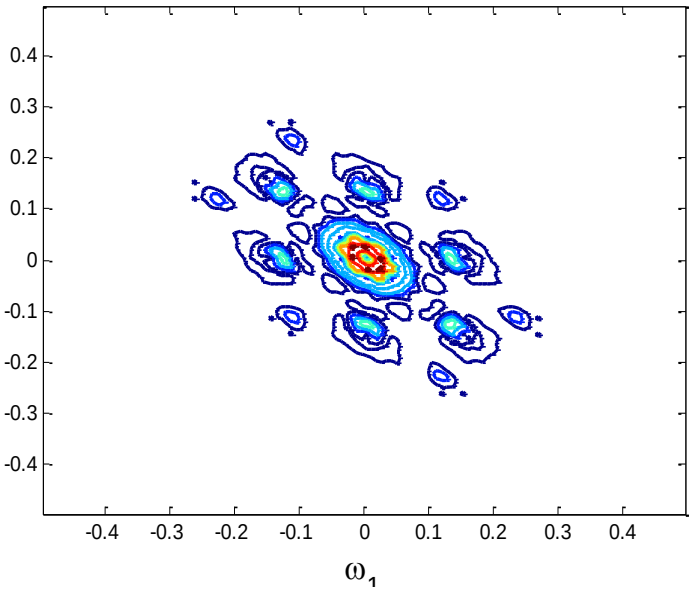
- More specifically, to capture nonlinearities due to MSM processing chain, we focus *on intermodulation distortion-spread*.

Bicoh. Mag. Spec. the Original & Cloned Recordings

$|C_3^x(\omega_1,\omega_2)|$ of Direct Audio Recording (Mic 2)



$|C_3^x(\omega_1,\omega_2)|$ of the Audio Recording Subjected to RA



Feature Extraction

- It can be observed that bicoherence magnitude plots exhibit 12 regions of symmetry.
- Image moments are commonly used to characterize images based on scale, centroid, and orientation.
- The *scale and rotation invariant Hu moments* are used to detect nonlinearities due to cloning attack.

Feature Extraction: *Scale Invariant Hu Moments*

- The scale invariant Hu moments $\eta_{i,j}$ where i + j ≥ 2 is computed as,

$$\eta_{i,j} = \frac{\mu_{i,j}}{\mu_{0,0}^{\left(1+\frac{i+j}{2}\right)}}$$

here, 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 the bicoherence magnitude spectrum, b (. , .)

Feature Extraction: *Rotation Invariant Hu Moments*

- The rotation invariant Hu moments $I_i,\ i \geq 1$ is computed as,

$$
\begin{aligned}
I_1 &= \eta_{2,0} + \eta_{0,2} \\
I_2 &= (\eta_{2,0} + \eta_{0,2})^2 + (2\eta_{1,1})^2 \\
I_3 &= (\eta_{3,0} - 3\eta_{1,2})^2 + (3\eta_{1,2} - \eta_{0,3})^2 \\
I_4 &= (\eta_{3,0} + 3\eta_{1,2})^2 + (3\eta_{1,2} + \eta_{0,3})^2
\end{aligned}
$$

Cloned Recording Identification

- Distance-based similarity measure is used to distinguish between the direct and the cloned recordings.

  – More specifically, the Euclidian distance between the 12-D feature vector consisting of *invariant Hu moments* computed from the bicoherence magnitude spectra estimated from the direct and cloned recordings.

$$d_{R,C} = \|F_R - F_C\|_{L_2}$$

where

$$F = \{\mu_{2,0}, \mu_{3,0}, \cdots, \mu_{8,0}, I_1, I_2, I_3, I_4\}$$

Data Set

- A set of recordings of a male speaker (talker) reading a short sentence were made using three different built-in Laptop microphones.
- All original and cloned recordings were saved on a Laptop computers running Windows Vista OS.
- All original and cloned recordings were made with the following settings:
  - 1) mono-channel,
  - 2) 44100 Hz sampling frequency, and
  - 3) 16-bit resolution.
- For these recordings, the speaker was standing approximately 1.0 meter from the microphones places on a table.
- To simulate a replay attack (RA), each recording was cloned by playing an audio recording through a commercial grade speaker and capturing it using the same (used originally) microphone.

List of the Microphones Used

Mic. ID	Microphone Details
Mic. 1	Sony Viao
Mic. 2	Lenovo Thinkpad X60
Mic. 3	Dell Latitude

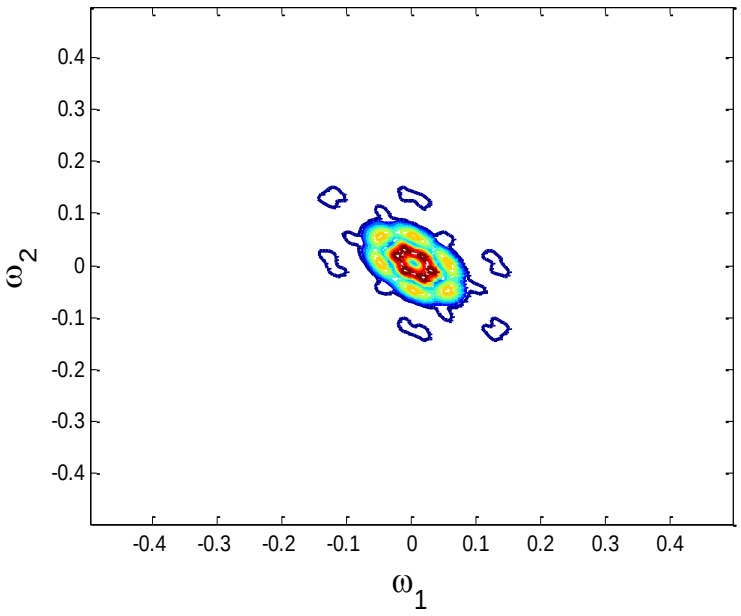
Parameter Settings Bicoh. Estimation

- Bicoherence is estimated from each audio segment using the direct (fft-based) approach (Nikias & Petropulu, 1993).

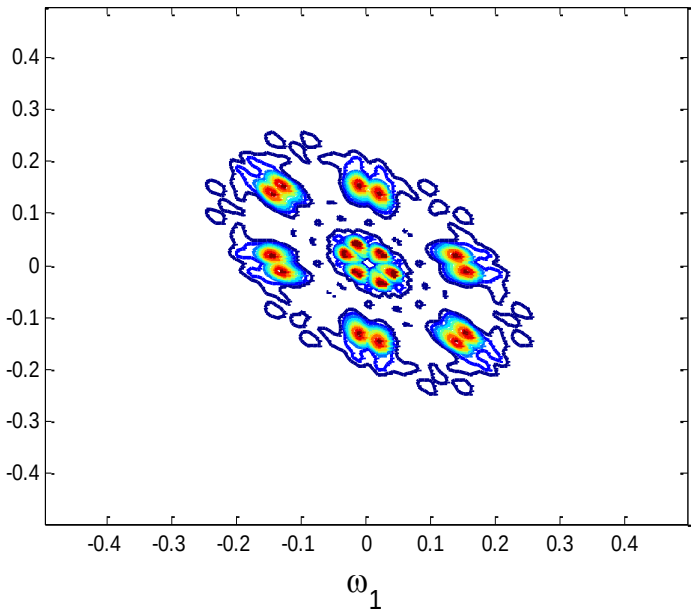
- The bicoherence is estimated with the following parameter settings:
  1. 64-point segment length,
  2. 128-point FFT length,
  3. no overlap, and
  4. 64-point Hamming window for time-domain smoothing.

Bicoh. Mag. Spec. the Original & Cloned Recordings

$|C_3^x(\omega_1,\omega_2)|$ of Direct Audio Recording (Mic 1)

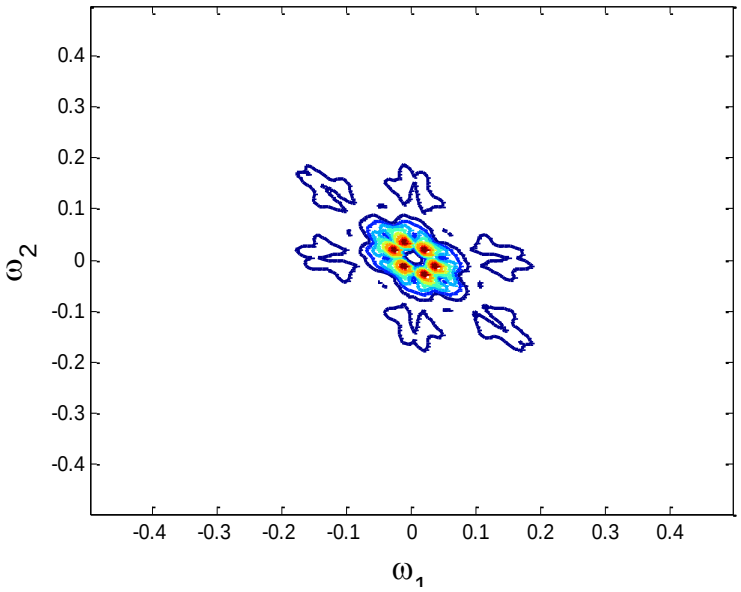


$|C_3^x(\omega_1,\omega_2)|$ of the Audio Recording Subjected to RA

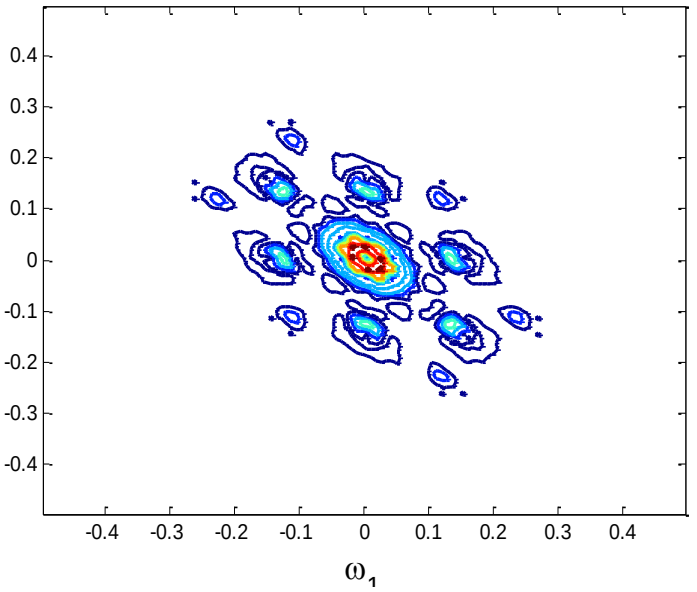


Bicoh. Mag. Spec. the Original & Cloned Recordings

$|C_3^x(\omega_1,\omega_2)|$ of Direct Audio Recording (Mic 2)

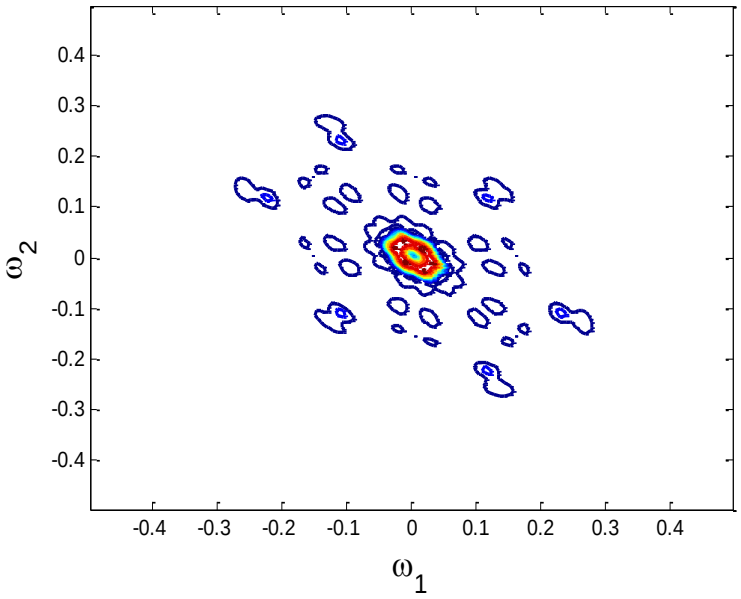


$|C_3^x(\omega_1,\omega_2)|$ of the Audio Recording Subjected to RA

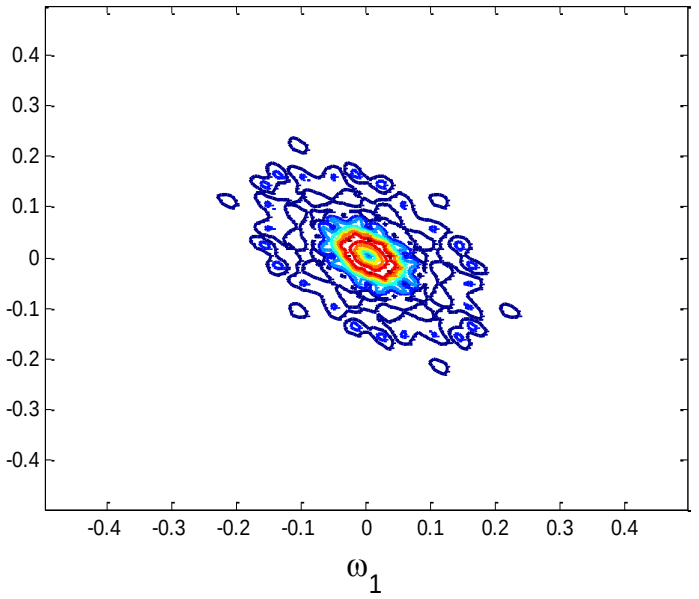


Bicoh. Mag. Spec. the Original & Cloned Recordings

$|C_3^x(\omega_1,\omega_2)|$ of Direct Audio Recording (Mic 3)



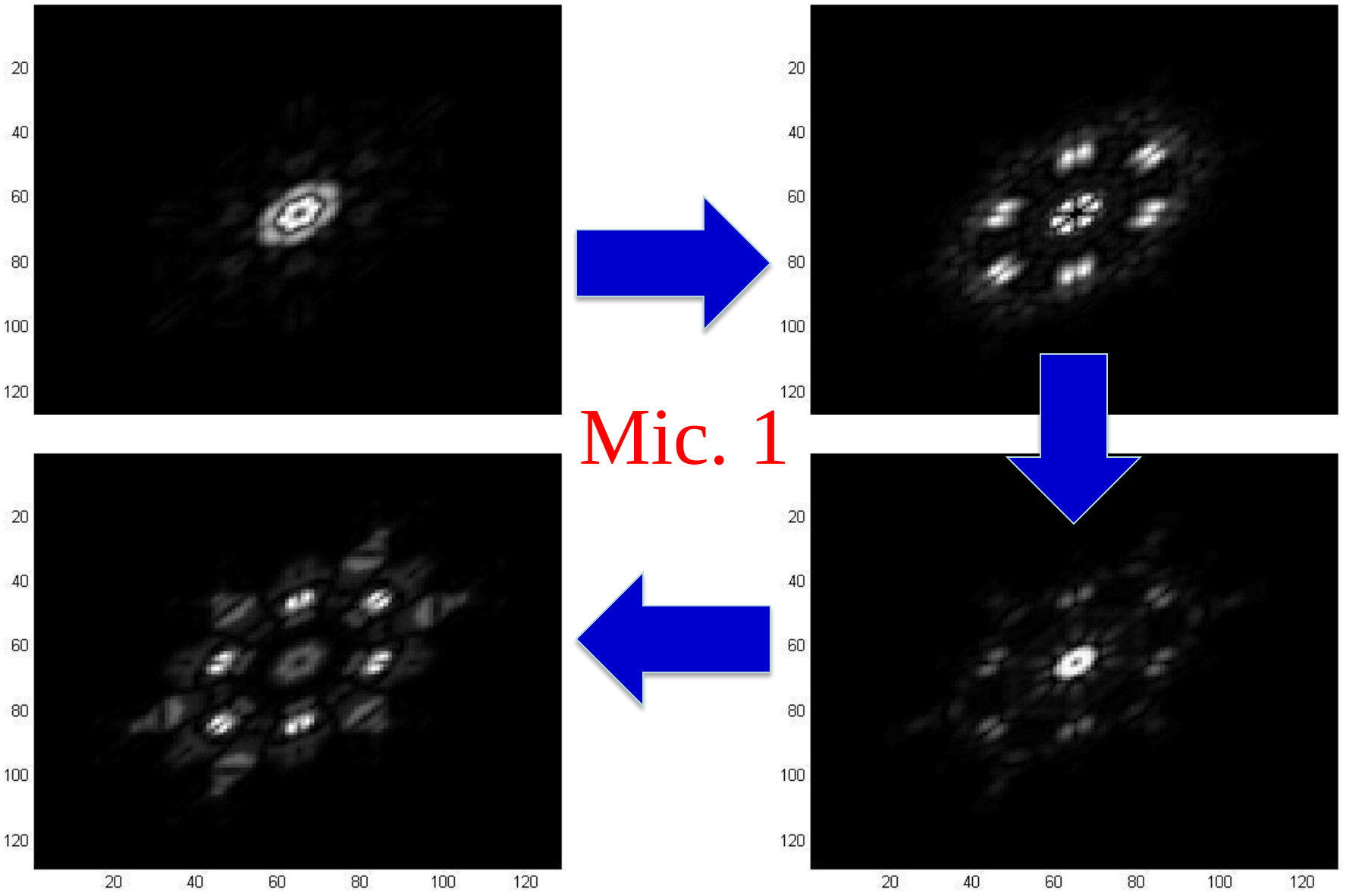
$|C_3^x(\omega_1,\omega_2)|$ of the Audio Recording Subjected to RA



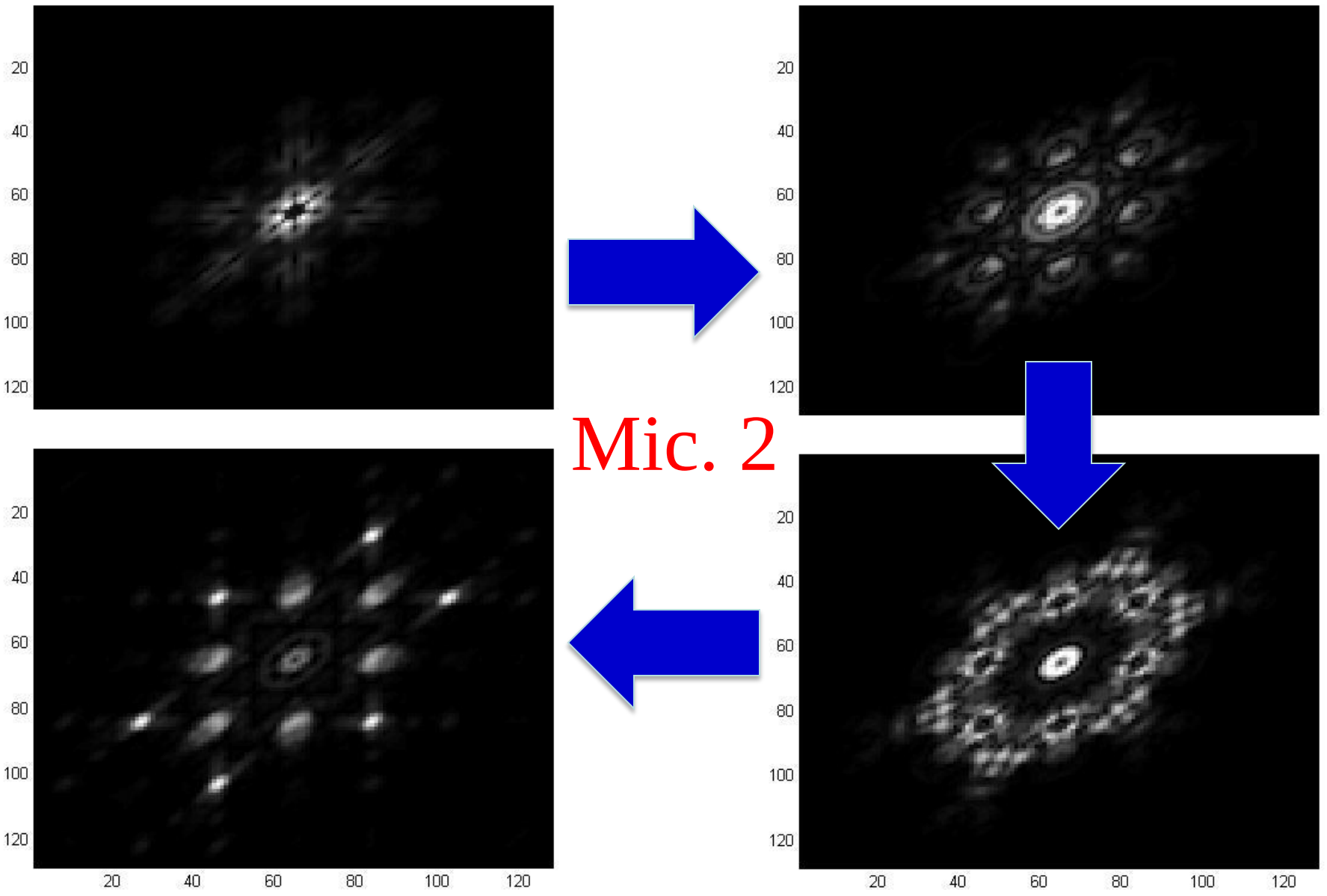
Cloning Recording Identification

Mic ID	Euclidian Distance between Original & Cloned Recordings	
	without Enhancement	*with Enhancement*
	D_{woE} $(\times 10^{11})$	D_{wE} $(\times 10^{11})$
Mic 1	2.9	5.8
Mic 2	1.9	12.3
Mic 3	5.1	18.8

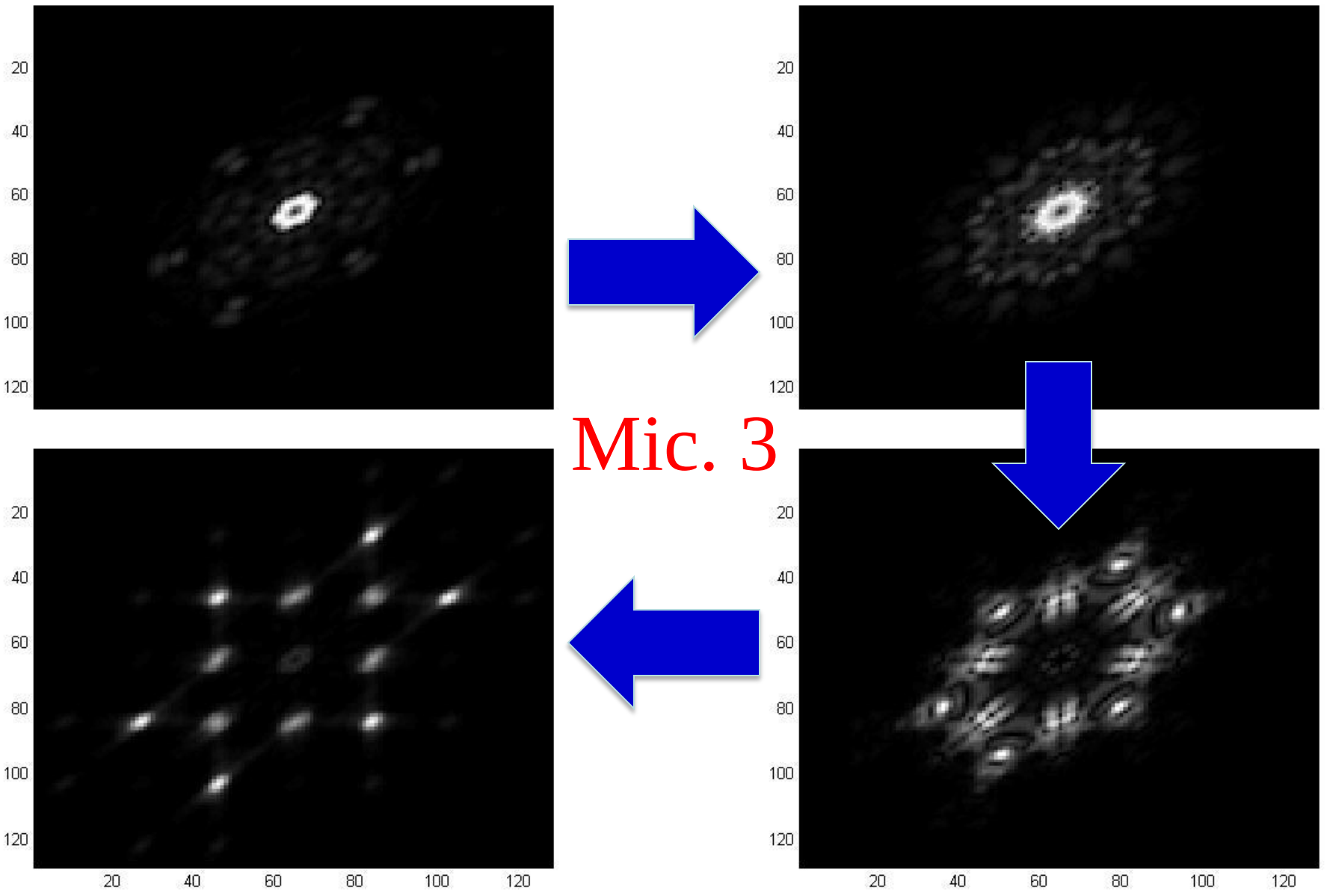
Bicoh. Mag. Spec. of 1st … 3rd –order RA



Bicoh. Mag. Spec. of 1st … 3rd –order RA



Bicoh. Mag. Spec. of 1st … 3rd –order RA



Conclusion

- Audio artifacts due to RA/cloning can be modeled and estimated.
- The proposed framework is applicable in securing speaker verification systems against RA.

Future Directions

- Investigate performance of existing speaker verification systems against RA.
- Investigate robustness against lossy compression, transcoding, additive noise attacks
- Integration of the proposed method with the existing speaker verification systems

Questions/Comments?