A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case

Zhizheng Wu[1,2], *Tomi Kinnunen*[3], *Eng Siong Chng*[1,2], *Haizhou Li*[1,2,4,5], *Eliathamby Ambikairajah*[5]

1. School of Computer Engineering, Nanyang Technological University, Singapore

2. Temasek Laboratories@NTU, Nanyang Technological University, Singapore

3. School of Computing, University of Eastern Finland, Finland

4. Human Language Technology Department, Institute for Infocomm Research, Singapore

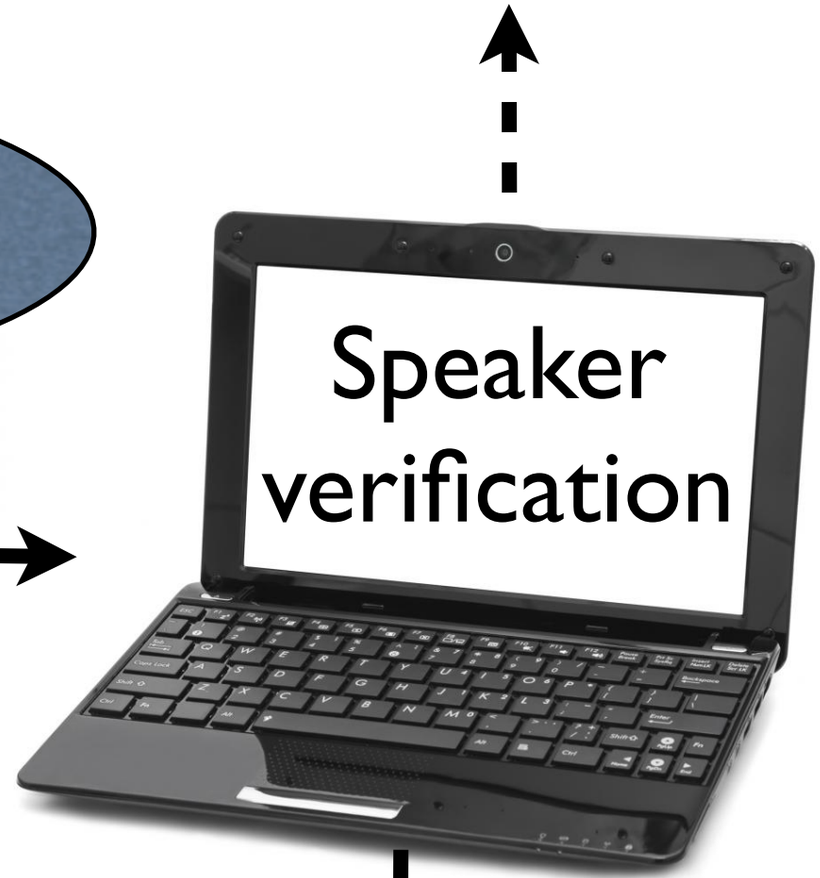
5. School of Electrical Engineering and Telecomm, University of New South Wales, Australia

Outline

- Introduction of speaker verification
- Spoofing attack
- Motivation
- Anti-spoofing techniques
- Experiments and results
- Conclusions and future work

Speaker verification

I am Obama, verify me!



Not Obama!

Spoofing attack in speaker verification

Voice conversion techniques can break a speaker verification system!

Motivation

- To prevent spoofing attack to speaker verification systems, we need to develop a technique to detect synthesized/converted speech

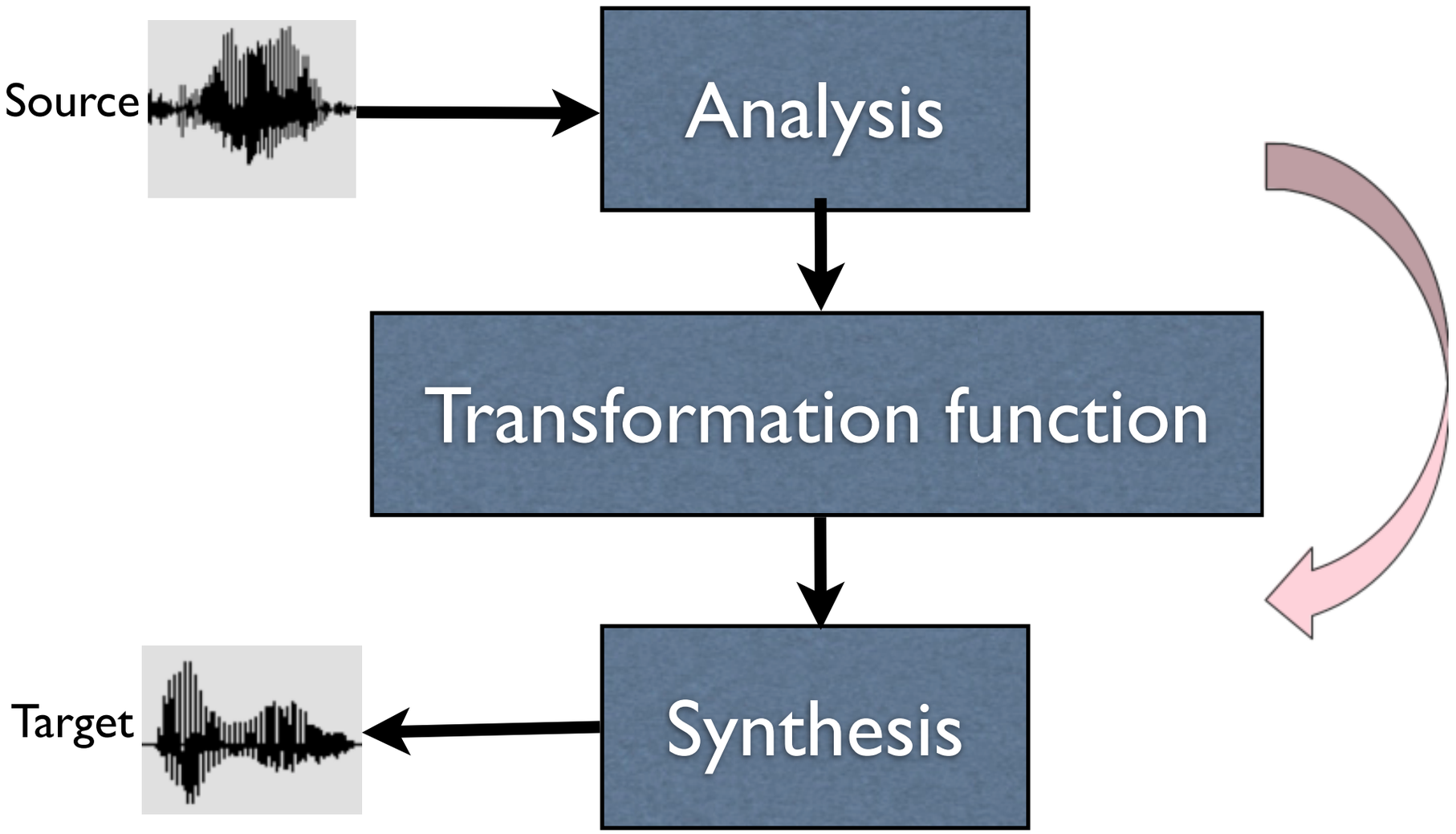
- Phase information was not given much attentions in speaker verification studies, but phase artifacts in synthesized speech is an informative cue

  - Zhizheng Wu, Eng Siong Chng, Haizhou Li, "*Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition*", Interspeech 2012.

  - Tomi Kinnunen, Zhizheng Wu, Kong Aik Lee, Filip Sedlak, Eng Siong Chng, Haizhou Li, "*Vulnerability of Speaker Verification Systems Against Voice Conversion Spoofing Attacks: the Case of Telephone Speech*", ICASSP 2012.

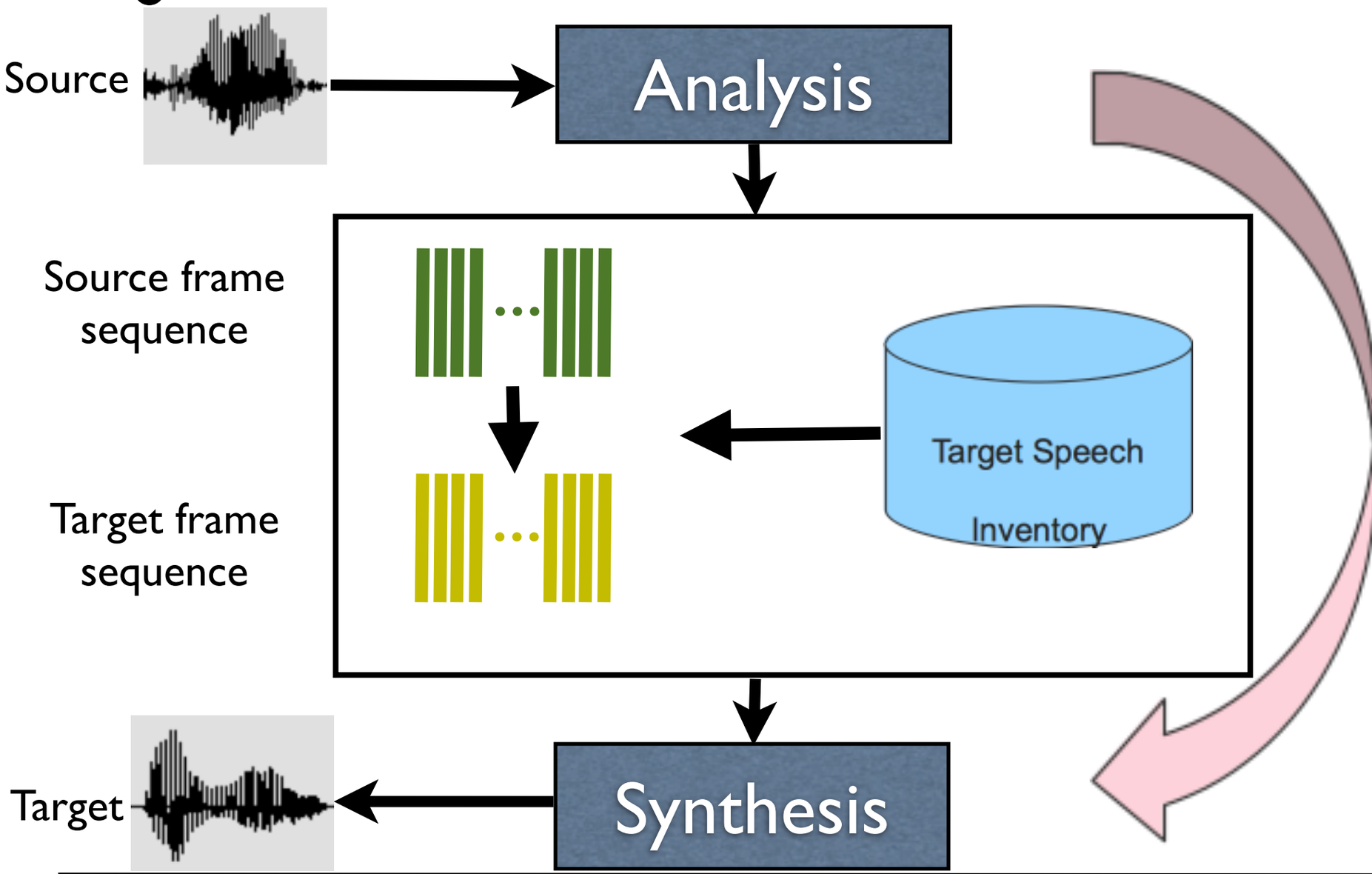
Overview of voice conversion (1/4)

- **GMM-based voice conversion**



Overview of voice conversion (2/4)

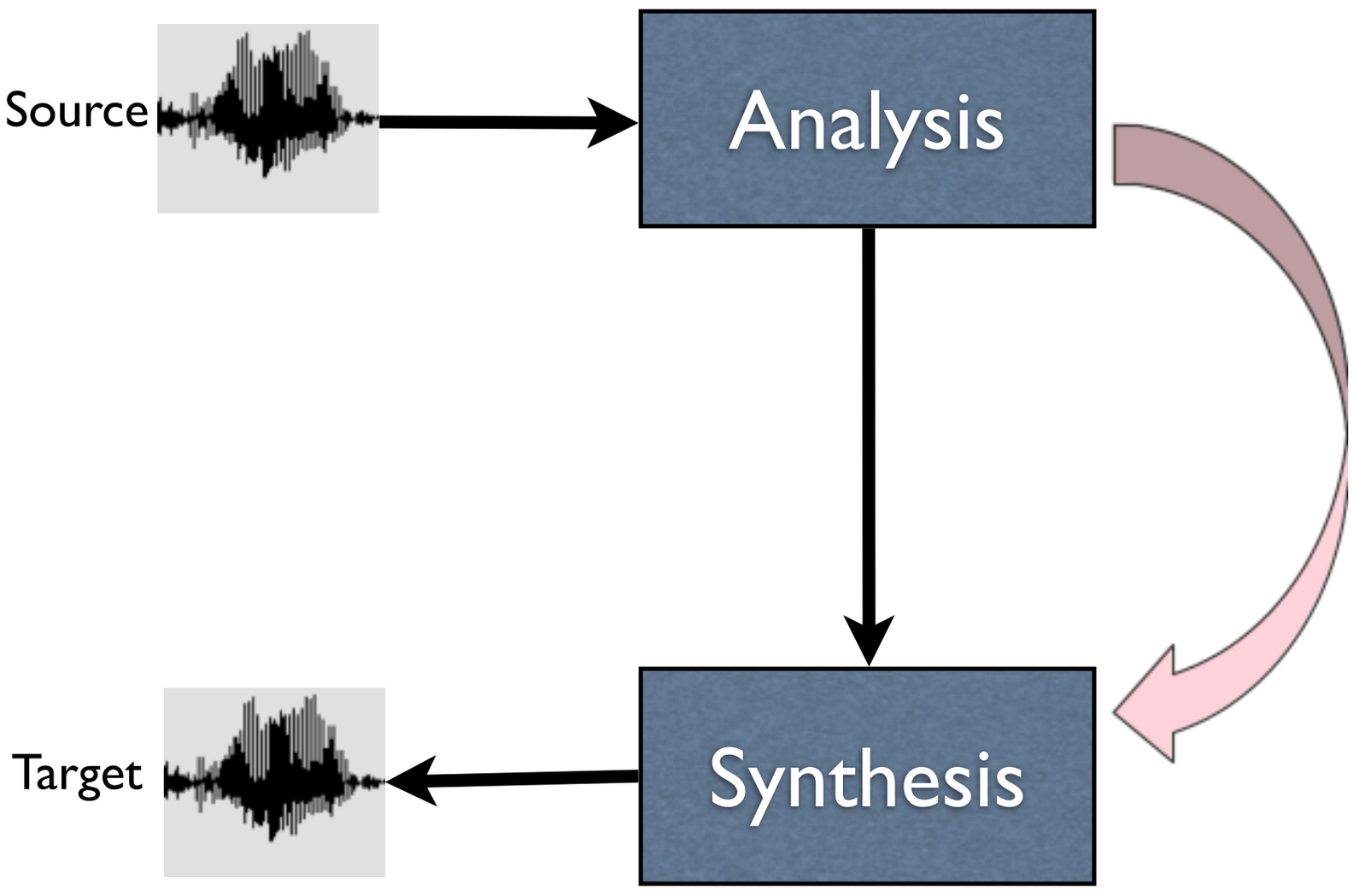
- **Unit-selection based voice conversion**



Phase artifacts created between analysis and synthesis

Overview of voice conversion (3/4)

- An analysis-synthesis pass-through without transformation



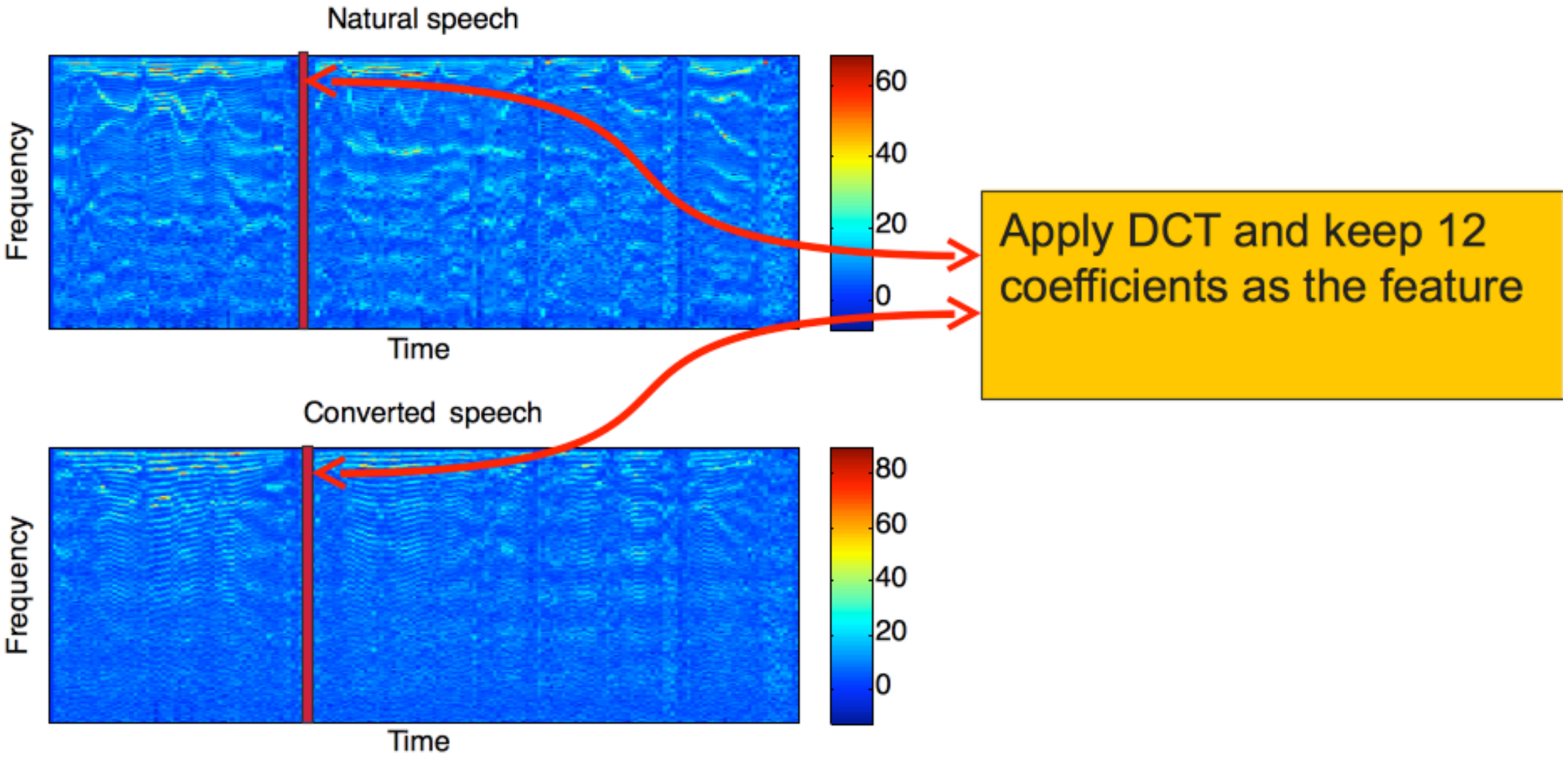
Phase artifacts created between analysis and synthesis

Overview of voice conversion (4/4)

- The three voice conversion systems all adopt analysis-synthesis module.
  - Use analysis module to extract features
  - Use synthesis filter to reconstruct speech signal from features
  - Hence, we can use the pass-through speech as training data for the synthetic speech detector

Anti-spoofing attack in speaker verification(1/3)

- **Feature for detection: modified group delay phase**



Anti-spoofing attack in speaker verification(2/3)

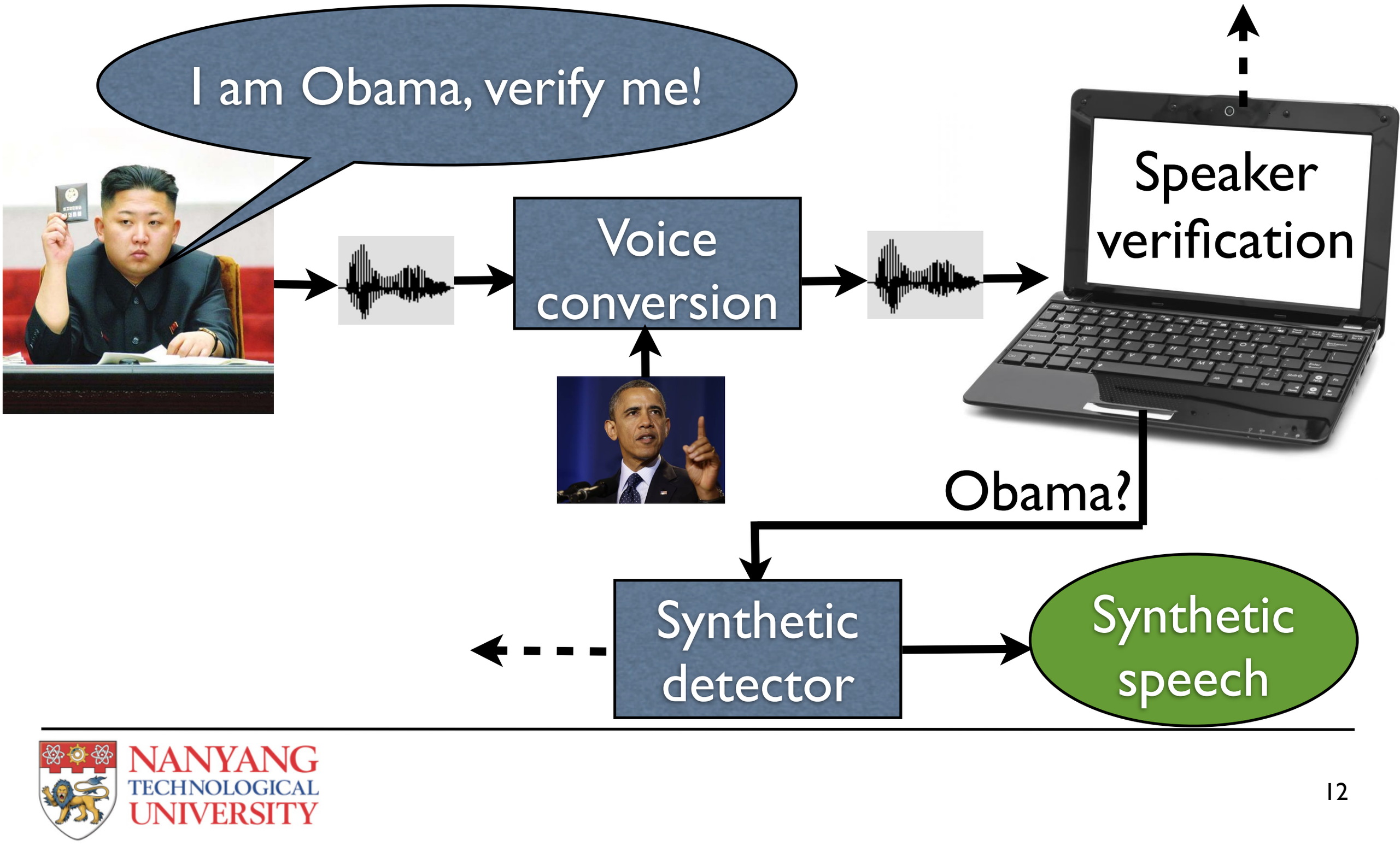
- **GMM-based detector**
  - **The decision is made based on the log-likelihood threshold**

$$\Lambda(C) = \log p(C|\lambda_{converted}) - \log p(C|\lambda_{natural})$$

  - **C is the feature vector sequence of a testing speech**
  - **$\lambda_{converted}$ is the GMM model for converted speech**
  - **$\lambda_{natural}$ is the GMM model for natural speech**
  - **512 Gaussian components are employed in this study**

NANYANG TECHNOLOGICAL UNIVERSITY

Anti-spoofing attack in speaker verification(3/3)



Experimental setups (1/3)

- **Original Dataset**
  - **subset of NIST 2006 SRE core task**

	Female	Male	Total
Unique speakers	298	206	504
Genuine trials	2, 349	1, 629	3, 978
Impostor trials	1, 636	1, 146	2, 782

  - **The duration of each conversation is about 5 minutes**

Experimental setups (2/3)

- **Spoofing dataset**
  - **Converted the 2, 782 impostor samples to the target speakers (the claimed identities)**
  - **Use *3conv4w* and *8conv4w* training sections for voice conversion function training**
  - **SPTK: http://sp-tk.sourceforge.net/ is used for analysis-synthesis**
    - **Analysis: Mel-cepstral analysis**
    - **Synthesis: MLSA filter**

Experimental setups (3/3)

- **Speaker verification systems**
  - **GMM Joint factor analysis (GMM-JFA) system**
    - **Models the intersession and speaker variability in the GMM supervector space**
  - **Probabilistic linear discriminant analysis system**
    - **PLDA system is similar as JFA system, but use i-vector as the basis for factor analysis**

Experimental results

- Performance of speaker verification systems with/without anti-spoofing attack
  - Equal error rate (EER)
    - The lower of EER, the better performance

Voice conversion	EER(%)			
	Without anti-spoofing		With anti-spoofing	
	GMM-JFA	PLDA	GMM-JFA	PLDA
Baseline (No conversion)	3.24	2.99	3.13	2.88
GMM conversion	17.36	19.29	0.0	0.0
Unit-selection conversion	32.54	41.25	1.64	1.71

Conclusions and future work

- **Conclusions**
  - **Voice conversion techniques present a threat to state-of-the-art speaker verification systems**
  - **Phase features are useful for detecting the synthetic speech from natural speech**
- **Future work:**
  - **Investigate vocoder independent features**
  - **Investigate temporal features for synthetic speech detection**