NANYANG
TECHNOLOGICAL
UNIVERSITY

# Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition

**Zhizheng Wu[1], Eng Siong Chng[1], Haizhou Li[1,2,3]**

*[1]School of Computer Engineering, Nanyang Technological University, Singapore*
*[2]Human Language Technology Department, Institute for Infocomm Research, Singapore*
*[3]School of EE & Telecom, University of New South Wales, Australia*

*12-Sep-2012*

# Outline

- Motivation
- Voice conversion overview
- Phase feature extraction
- Experiments
- Conclusions

# Motivation

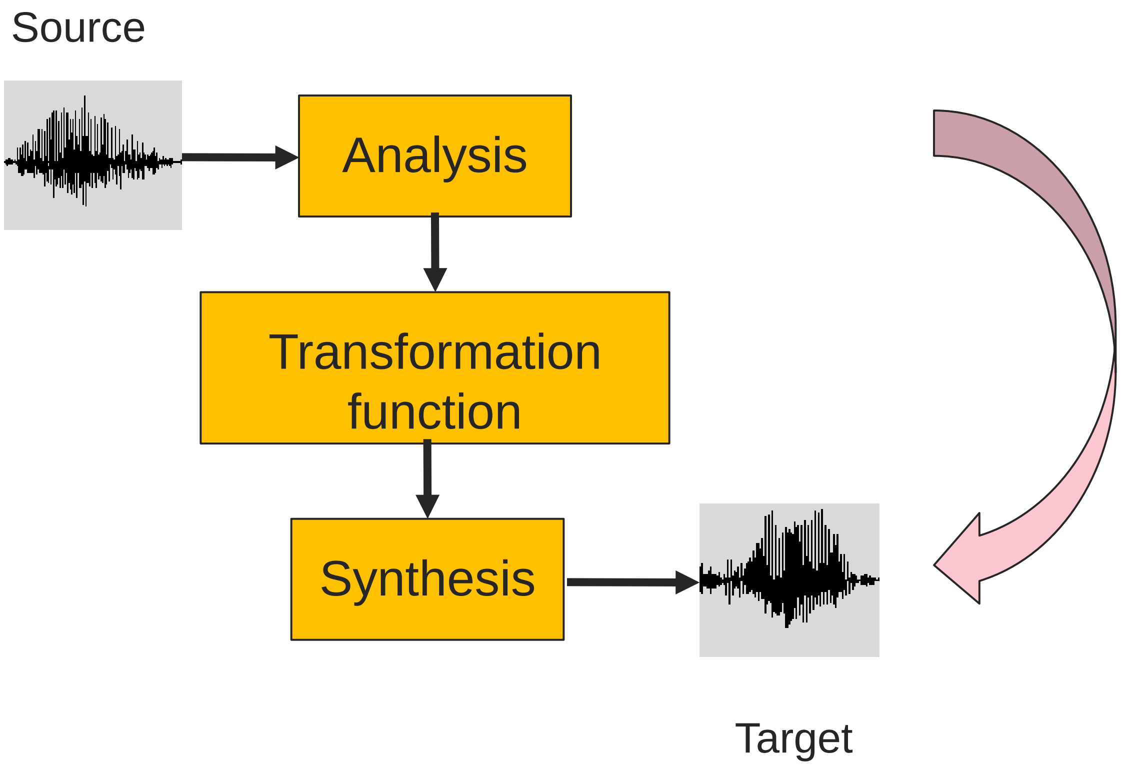
- We would like to detect converted speech (synthetic speech) to prevent spoofing attack against speaker verification system
- Phase artifacts in synthetic speech is an informative cue. We study the ways of phase feature extraction

1. 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.

2. Zhizheng Wu, Eng Siong Chng, Haizhou Li, "Speaker verification system against two different voice conversion techniques in spoofing attacks", Technical Report (http://www3.ntu.edu.sg/home/wuzz/), 2012.

# Overview of Voice Conversion（1/3）

- GMM-based voice conversion



Source

Analysis

Transformation function

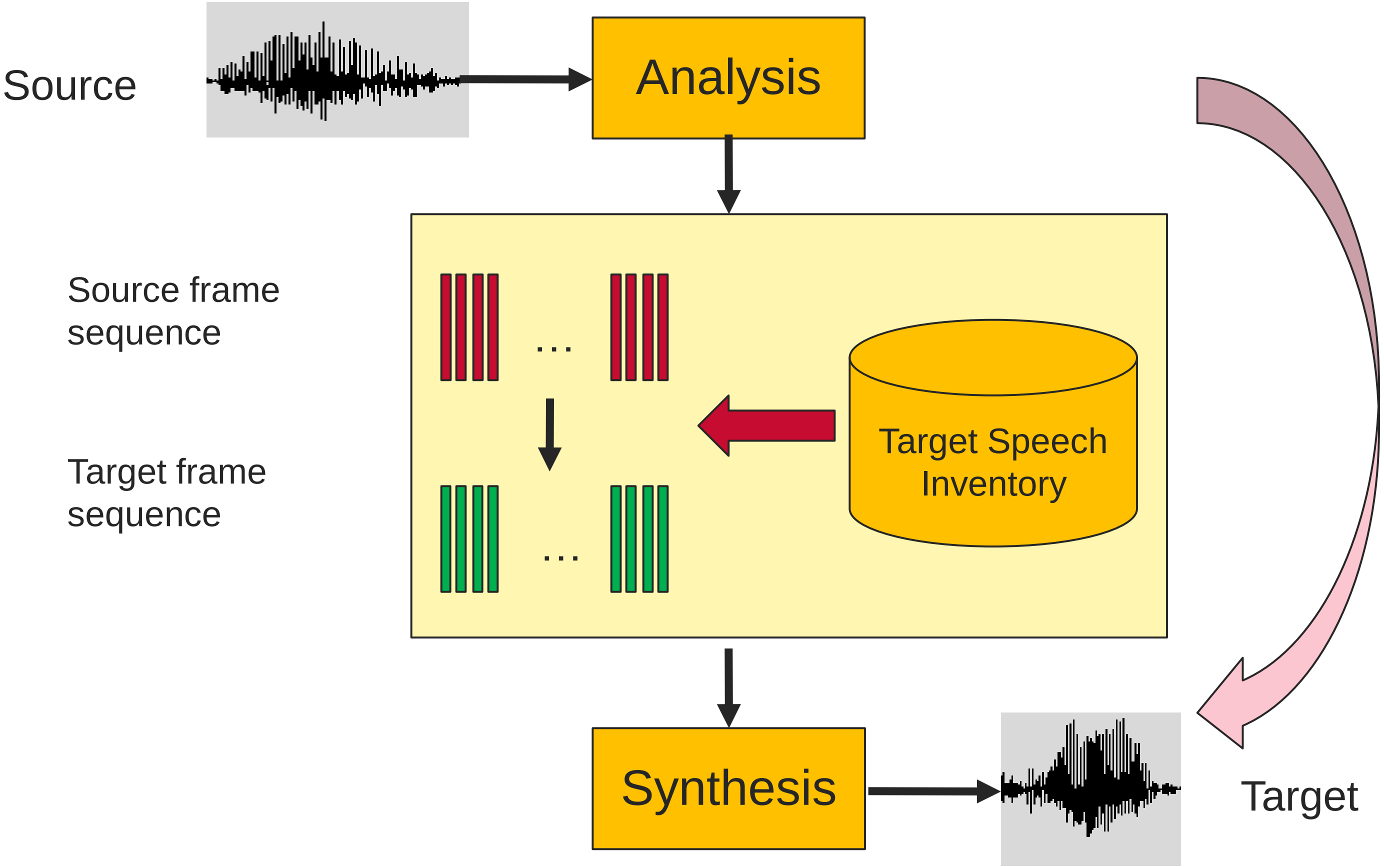
Synthesis

Target

Phase artifacts created between analysis and synthesis !

# Overview of Voice Conversion（2/3）

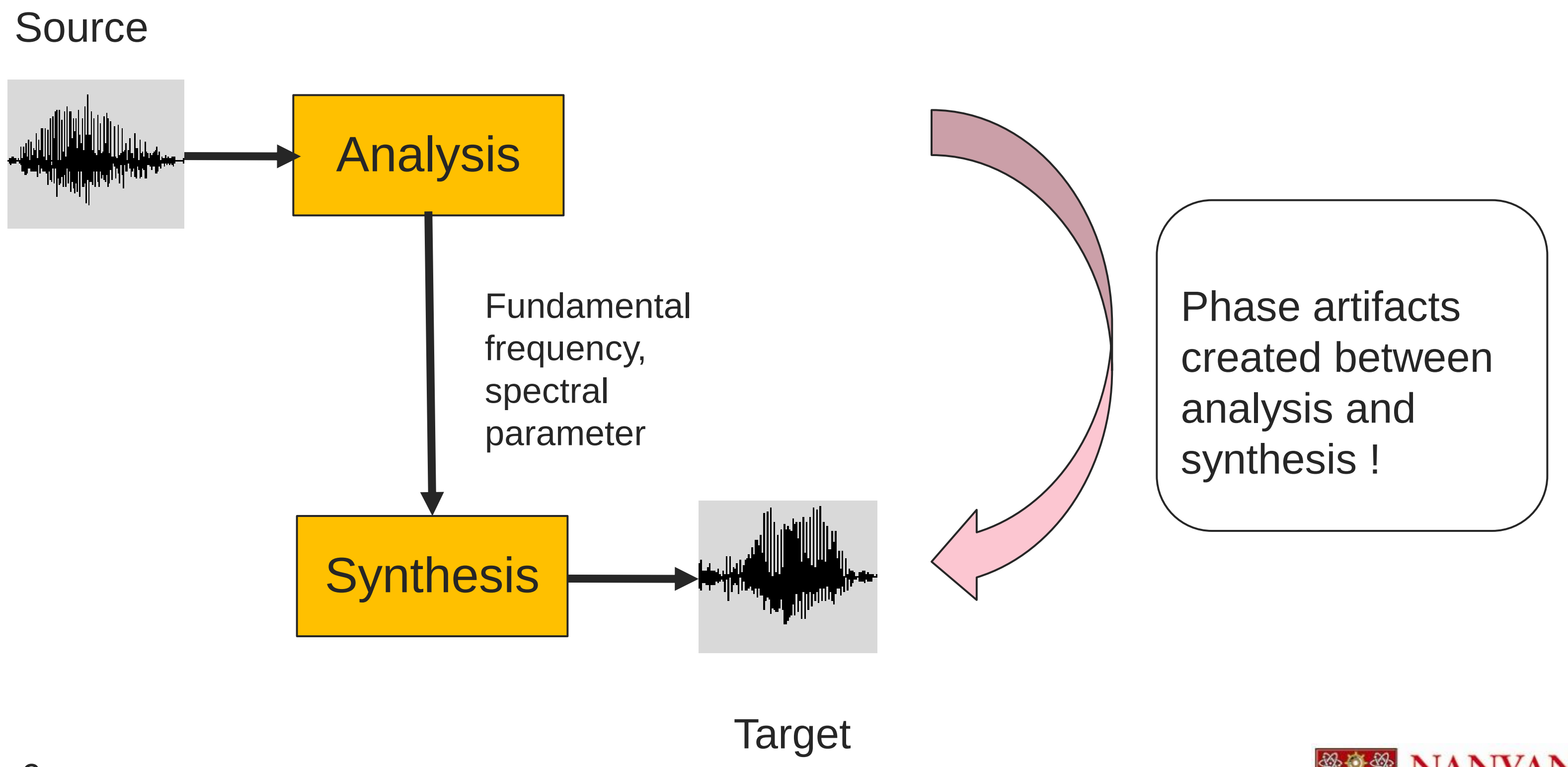
- Unit-selection based voice conversion



Phase artifacts created between analysis and synthesis !

# Overview of Voice Conversion（3/3）

- An analysis-synthesis *pass-through* without transformation



Source

Analysis

Fundamental frequency, spectral parameter

Synthesis

Target

Phase artifacts created between analysis and synthesis !

# Phase Artifacts

- Voice conversion techniques focus on spectral conversion
  - Magnitude spectrum contains more information
  - Many vocoders usually use random phase, not the original phase to reconstruct the speech

K.K. Paliwal and L.D. Alsteris, “On the usefulness of STFT phase spectrum in human listening tests,” Speech Communication, vol. 45, no. 2, pp. 153–170, 2005.

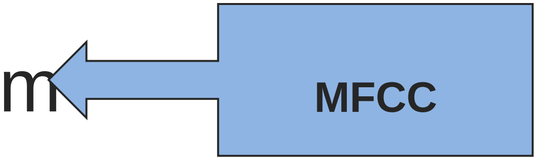
# Phase feature extraction

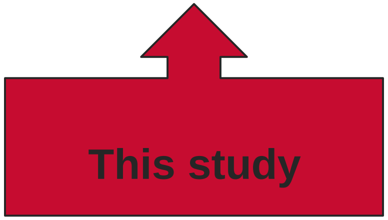
- Short-time Fourier transform of signal $x(n)$

$$X(w)=\left|X(w)\right|e^{jj(w)}$$

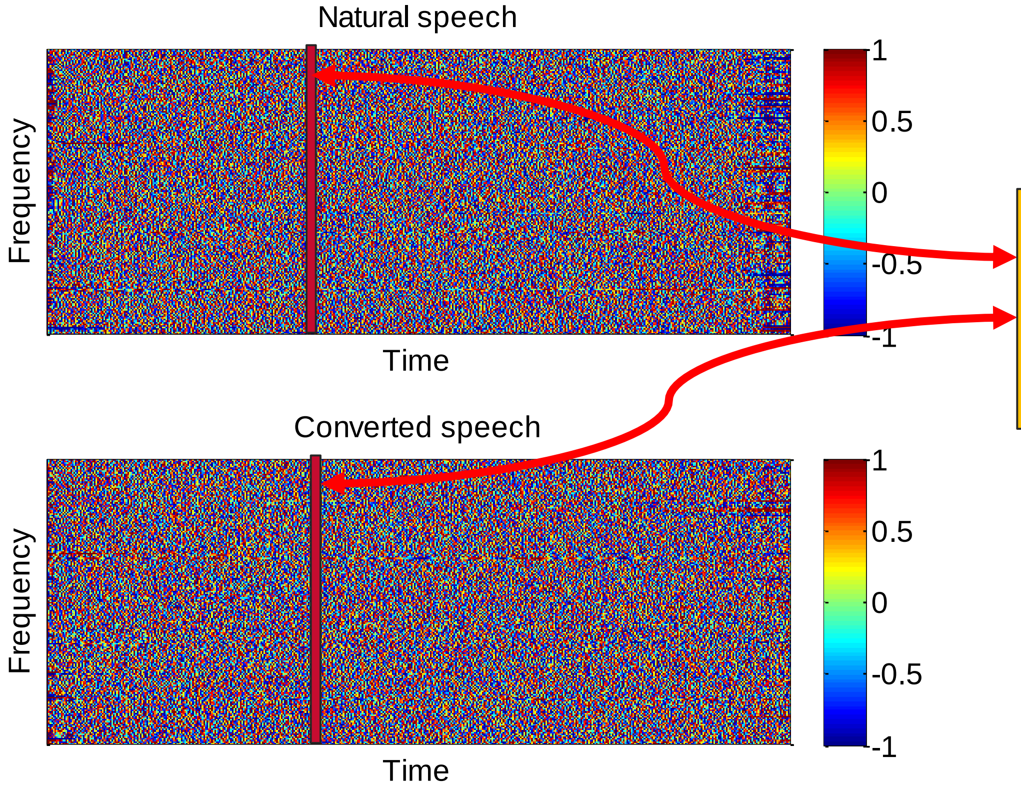
$\left|X(w)\right|$ is the magnitude spectrum

$j(w)$ is the phase spectrum


MFCC

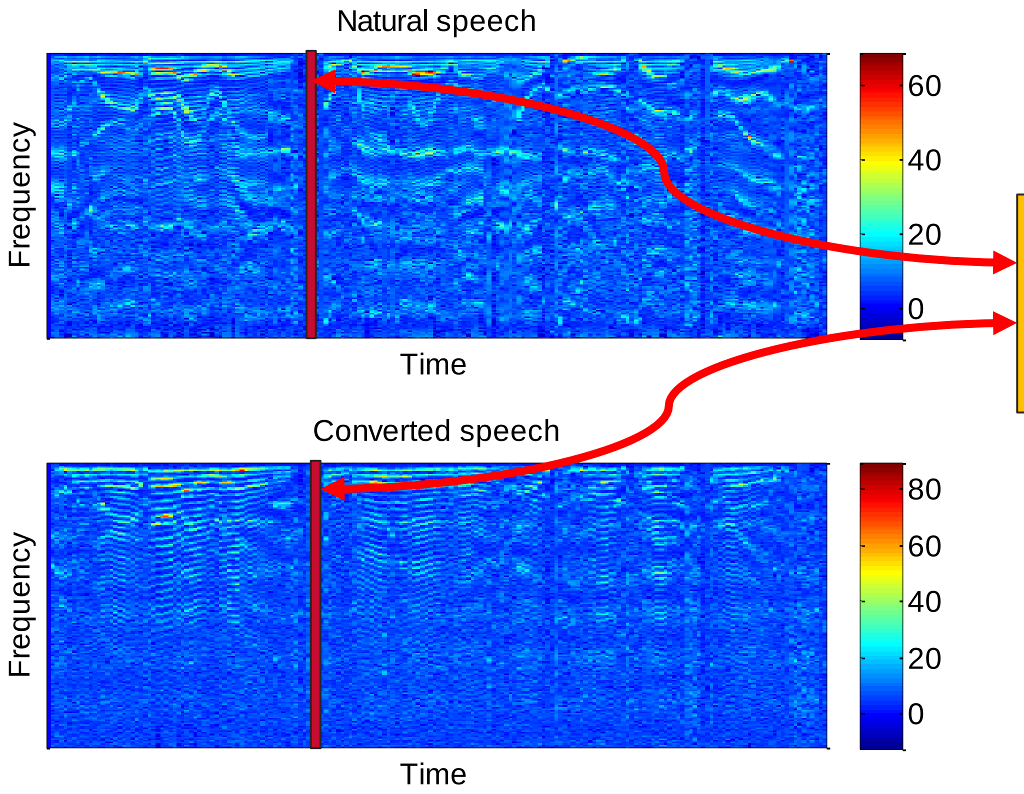

This study

# Cosine Normalized Phase Feature (*Cos-phase*)



Apply discrete cosine function (DCT) and keep 12 coefficients as the feature

# Modified group delay phase
## *(MGD-phase)*



Apply DCT and keep 12 coefficients as the feature

# Synthetic speech detector

- GMM-based detector

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

*C* is the feature vector sequence of a speech signal

$\lambda_{converted}$ is GMM model for converted speech

$\lambda_{natural}$ is GMM model for natural speech

We use 512 Gaussian components in this study.

# Experimental setups

- Corpus: a subset of NIST SRE 2006

| Training set (number of sessions) | |
|---|---|
| Natural model | Converted model |
| 100 | 100 |

  - The duration of each session is 5 minutes
  - Three training situations for converted model
    - GMM-based converted speech for training
    - Unit-selection based converted speech for training
    - *Pass-through* speech for training

We will conduct three experiments under the three training situations

# Experimental setups

| Testing set (number of sessions) | | |
|---|---|---|
| Natural | Converted | |
| | GMM | Unit-selection |
| 1, 500 | 1, 000 | 1, 000 |

- – Testing set: in total 3500 sessions.
- – Evaluation metric: Equal error rate
  - Natural to converted
  - Converted to natural

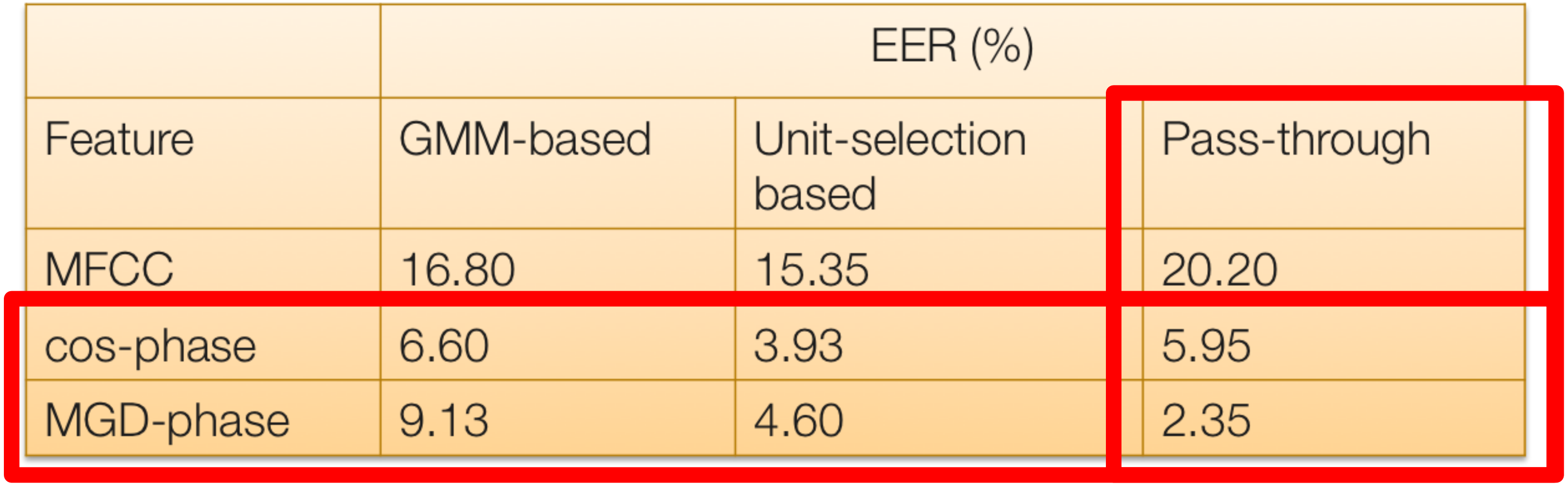
# Experimental setups

- Spoofing attack corpus construction
  - SPTK: http://sp-tk.sourceforge.net/
    - Analysis: Mel-cepstral analysis
    - Synthesis: MLSA filter

1. 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.
2. Zhizheng Wu, Eng Siong Chng, Haizhou Li, "Speaker verification system against two different voice conversion techniques in spoofing attacks", Technical Report (http://www3.ntu.edu.sg/home/wuzz/), 2012.

# Results:

- 3 speech models vs 3 features for synthetic speech detection



| | EER (%) | | |
|---|---|---|---|
| Feature | GMM-based | Unit-selection based | Pass-through |
| MFCC | 16.80 | 15.35 | 20.20 |
| cos-phase | 6.60 | 3.93 | 5.95 |
| MGD-phase | 9.13 | 4.60 | 2.35 |

# Conclusions

- Phase artifacts are useful in detecting the synthetic speech
- When transformation technique is *unknown*, we may use analysis-synthesis *pass-through* method to simulate converted data

# Thank you!