Voice Conversion and Spoofing Attack on Speaker Verification Systems

Haizhou Li

Institute for Infocomm Research (I^2R), Singapore

Acknowledgements: Zhizheng Wu, Eng Siong Chng, NTU Singapore

APSIPA ASC 2013

APSIPA Annual Summit and Conference

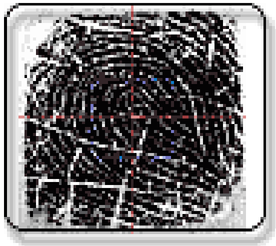
Kaohsiung, Taiwan. Oct. 29 - Nov. 1, 2013

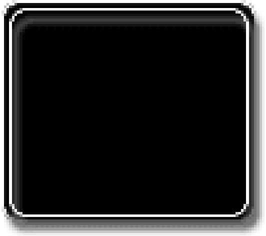
A*STAR I^2R

Outline

- Introduction
- Speaker verification
- Voice conversion and spoofing attack
- Anti-spoofing attack
- Future research

Introduction



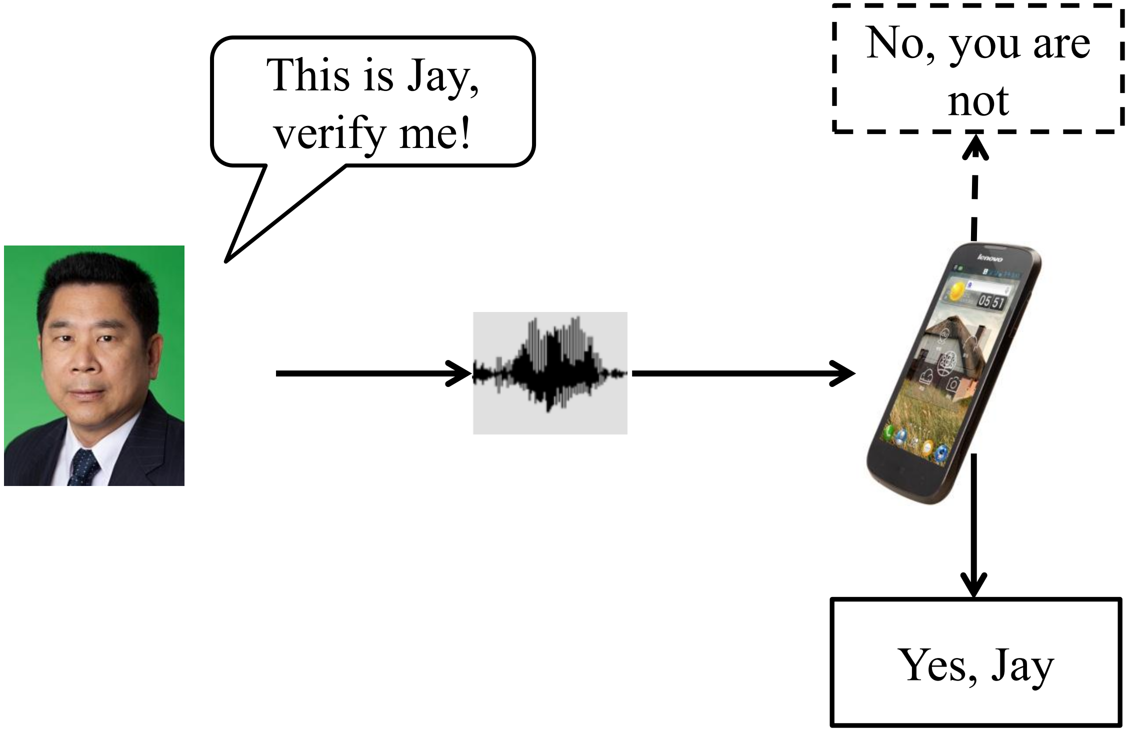


Authentication

To decide ‘Who you are’ based on ‘What you have’ and ‘What you know’

Biometrics

To verify identity of a living persons based on behavioral and physiological characteristics



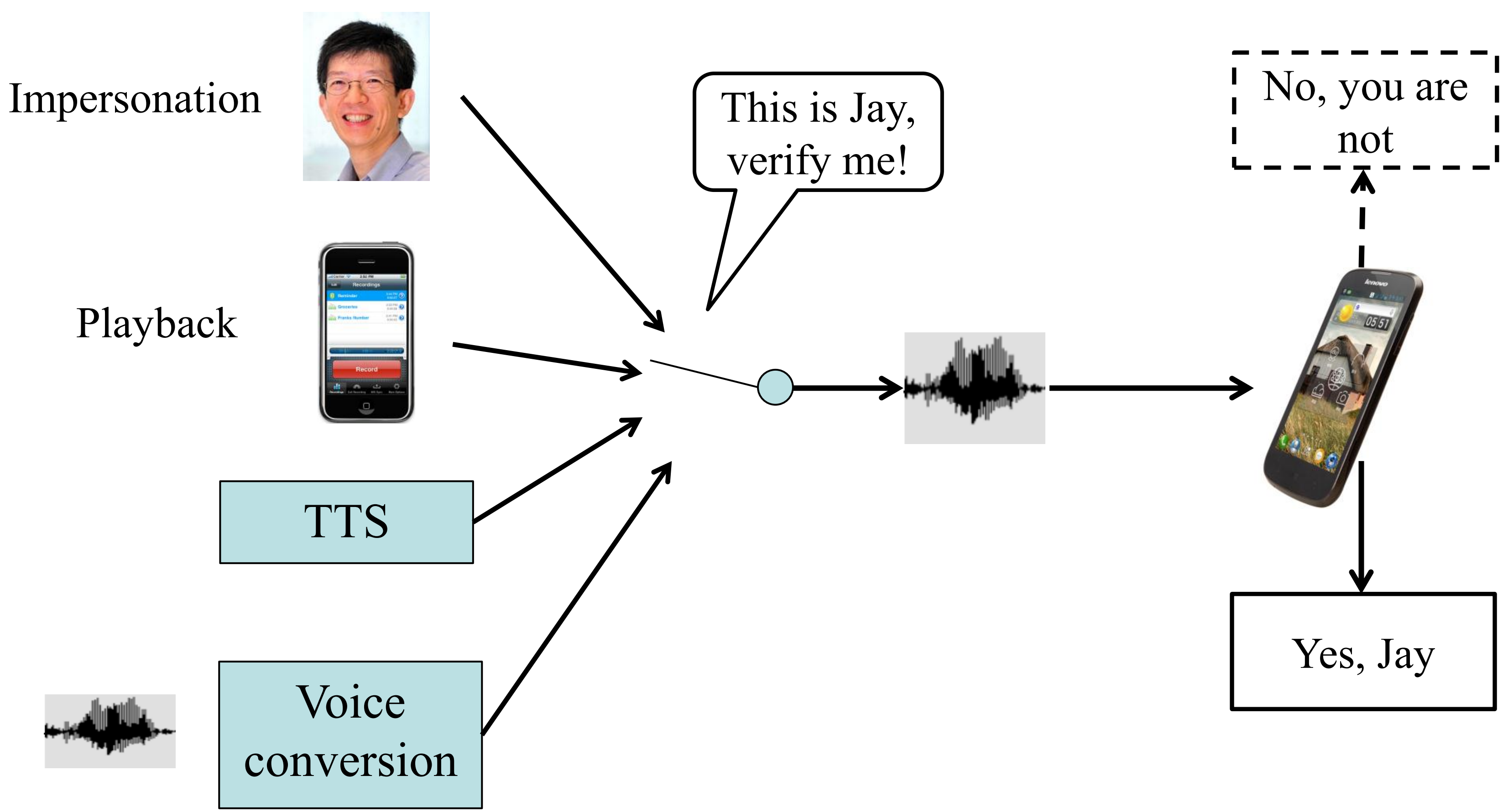
Mode

- Text-Dependent
- Text-Independent (Language-Independent)

A*STAR

Spoofing Attack | Speaker Recognition

Spoofing attack is to use a falsifying voice as the system input

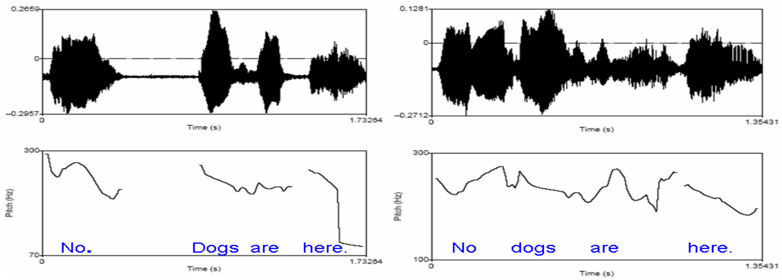


Summary of spoofing attack techniques

Spoofing technique	Accessibility (practicality)	Effectiveness (risk)	
		Text-independent	Text-dependent
Impersonation	Low	Low/unknown	Low/unknown
Playback	High	High	Low (promoted text) to high (fixed phrase)
Speech synthesis	Medium to High	High	High
Voice conversion	Medium to High	High	High

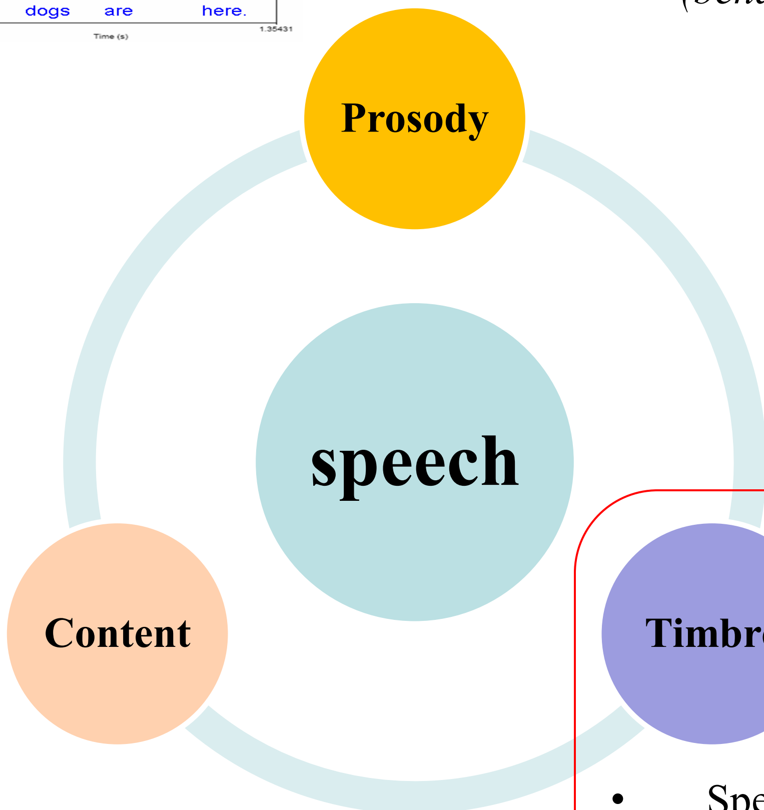
- Introduction
- Speaker verification
- Voice conversion and spoofing attack
- Anti-spoofing attack
- Future research

Speaker Verification

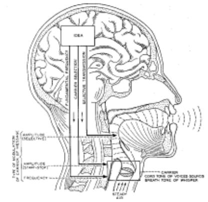


- Speech to Singing Synthesis
- Expressive Speech Synthesis

(behavioral characteristics)



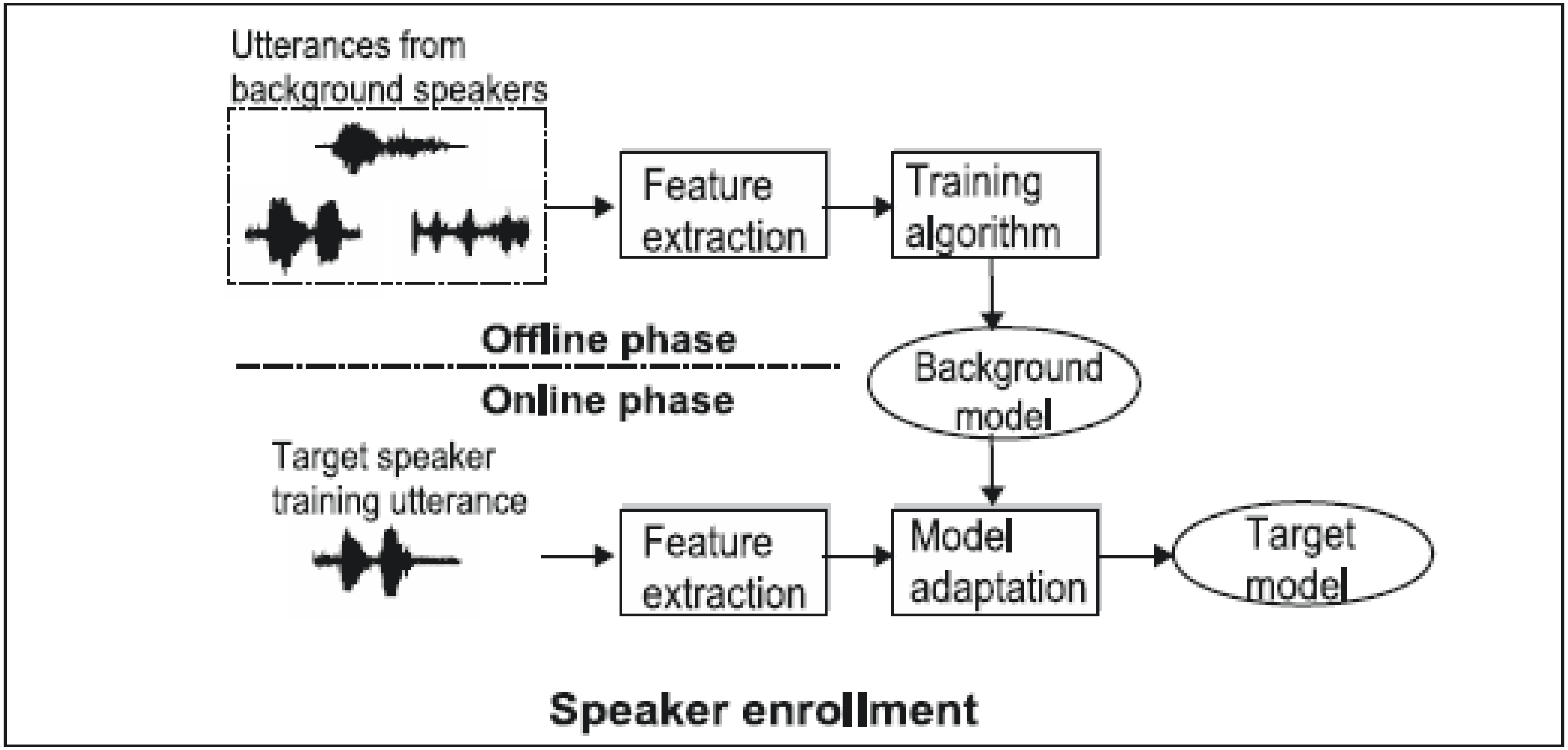
- Text-to-Speech
- Speech-to-Text



- Speaker Recognition
- Voice Conversion
- Voice Impersonation

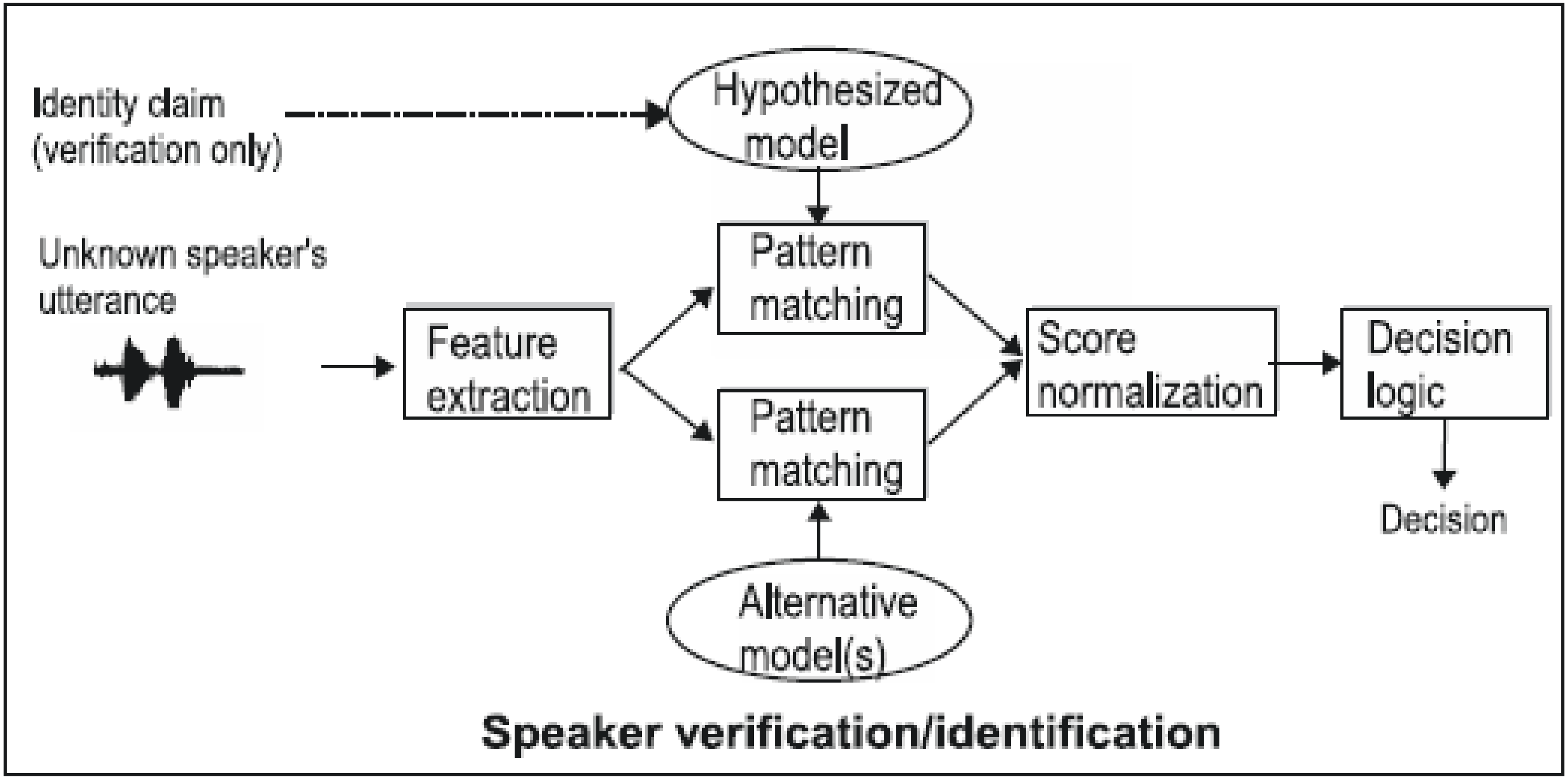
(physiological characteristics)

Speaker Verification



Tomi Kinnunen and Haizhou Li, “An Overview of Text-Independent Speaker Recognition: from Features to Supervectors”, Speech Communication 52(1): 12--40, January 2010

Speaker Verification



Tomi Kinnunen and Haizhou Li, “An Overview of Text-Independent Speaker Recognition: from Features to Supervectors”, Speech Communication 52(1): 12--40, January 2010

+ Robust against channel effects and noise

- Difficult to extract

- A lot of training data needed

- Delayed decision making

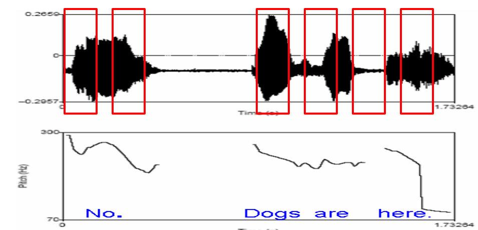
High-level features

Phones, idiolect (personal lexicon), semantics, accent, pronunciation

Learned (behavioral)
Socio-economic status, education, place of birth, language background, personality type, parental influence

Prosodic & spectro-temporal features

Pitch, energy, duration, rhythm, temporal features



+ Easy to extract

+ Small amount of data necessary

+ Text- and language independence

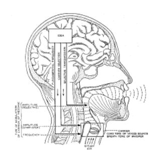
+ Real-time recognition

- Affected by noise and mismatch

Short-term spectral and voice source features

Spectrum, glottal pulse features

Physiological (organic)
Size of the vocal folds, length and dimensions of the vocal tract



Tomi Kinnunen and Haizhou Li, “An Overview of Text-Independent Speaker Recognition: from Features to Supervectors”, Speech Communication 52(1): 12--40, January 2010

Evaluation Metrics

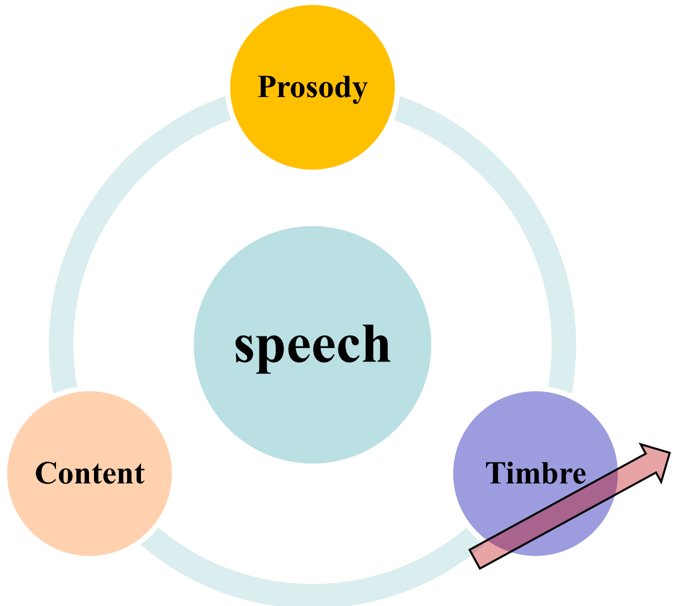
– Equal Error Rate (ERR): when *false alarm* equals *miss detection*
– Four categories of trial decisions in speaker verification

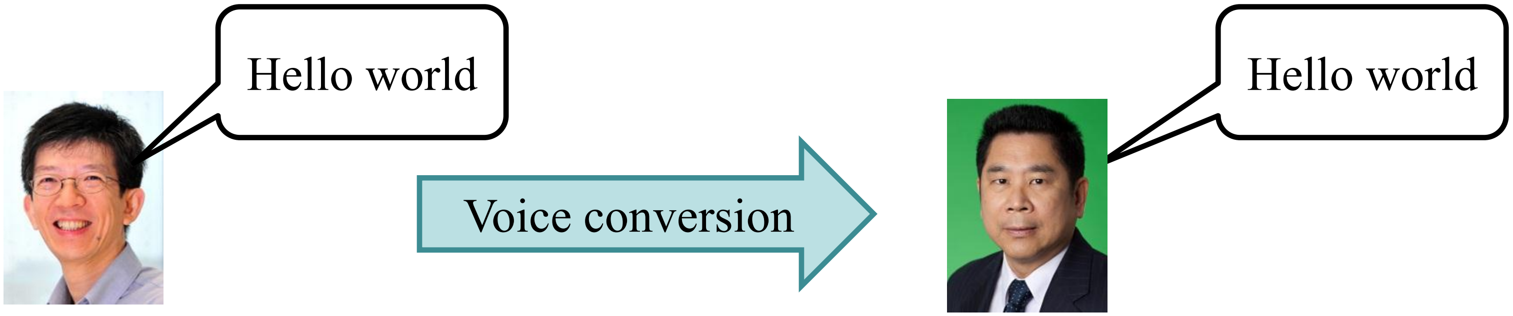
	Decision	
	Accept	Reject
Genuine	Correct acceptance	Miss detection
Impostor	False alarm (FAR)	Correct rejection

Some Observations

- Most systems use short-term spectral features (MFCC, LPCC) instead of segmental features (pitch contour, energy flow)
  - Systems sensitive to spectral features instead of prosodic features
  - Prosody could become a feature when detecting spoofing
- Most systems are sensitive to channels and noises
  - Same speaker, different channels/noises
  - Different speakers, same channel/noise
- All systems assume natural voice (genuine human voice) as inputs

- Introduction
- Speaker verification
- Voice conversion and spoofing attack
- Anti-spoofing attack
- Future research



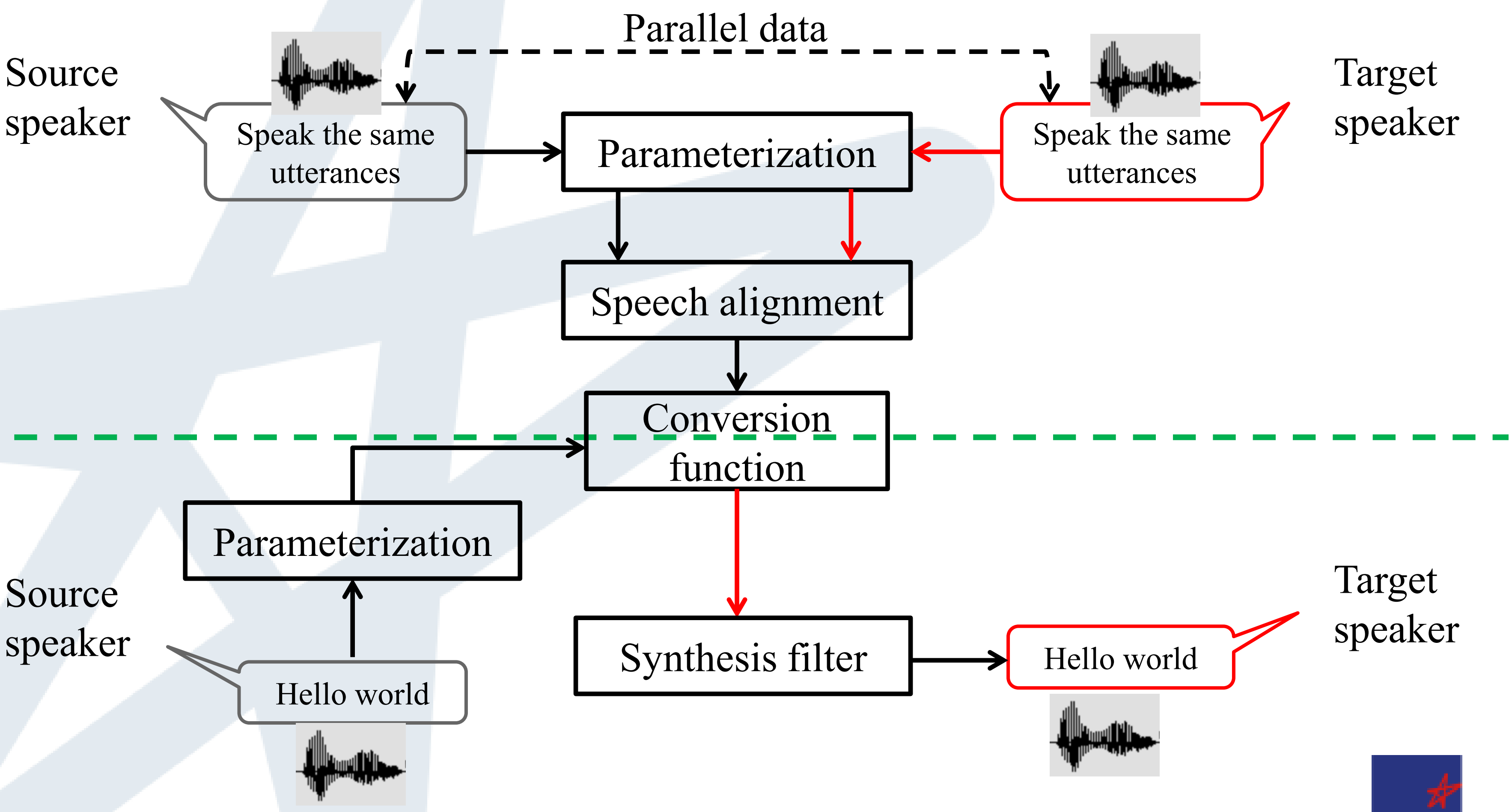


Source speaker's voice

Target speaker's voice

Yannis Stylianou, "Voice transformation: a survey." ICASSP 2009.

System Diagram



- Voice conversion demo
  - Using 10 utterances (around 30 seconds speech) to train the mapping function
  - Only transform the *timbre* while keeping the *prosody*

	Source	Target	Converted
Male-to-male			
Male-to-female			

- Four categories of trial decisions in speaker verification

	Decision	
	Accept	Reject
Genuine	Correct acceptance	Miss detection
Impostor	**False alarm (FAR)**	Correct rejection

- Spoofing attacks increase the false alarm, and thus increase equal error rate
- Move impostor's score distribution towards that of genuine

Voice Conversion Spoofing Attack

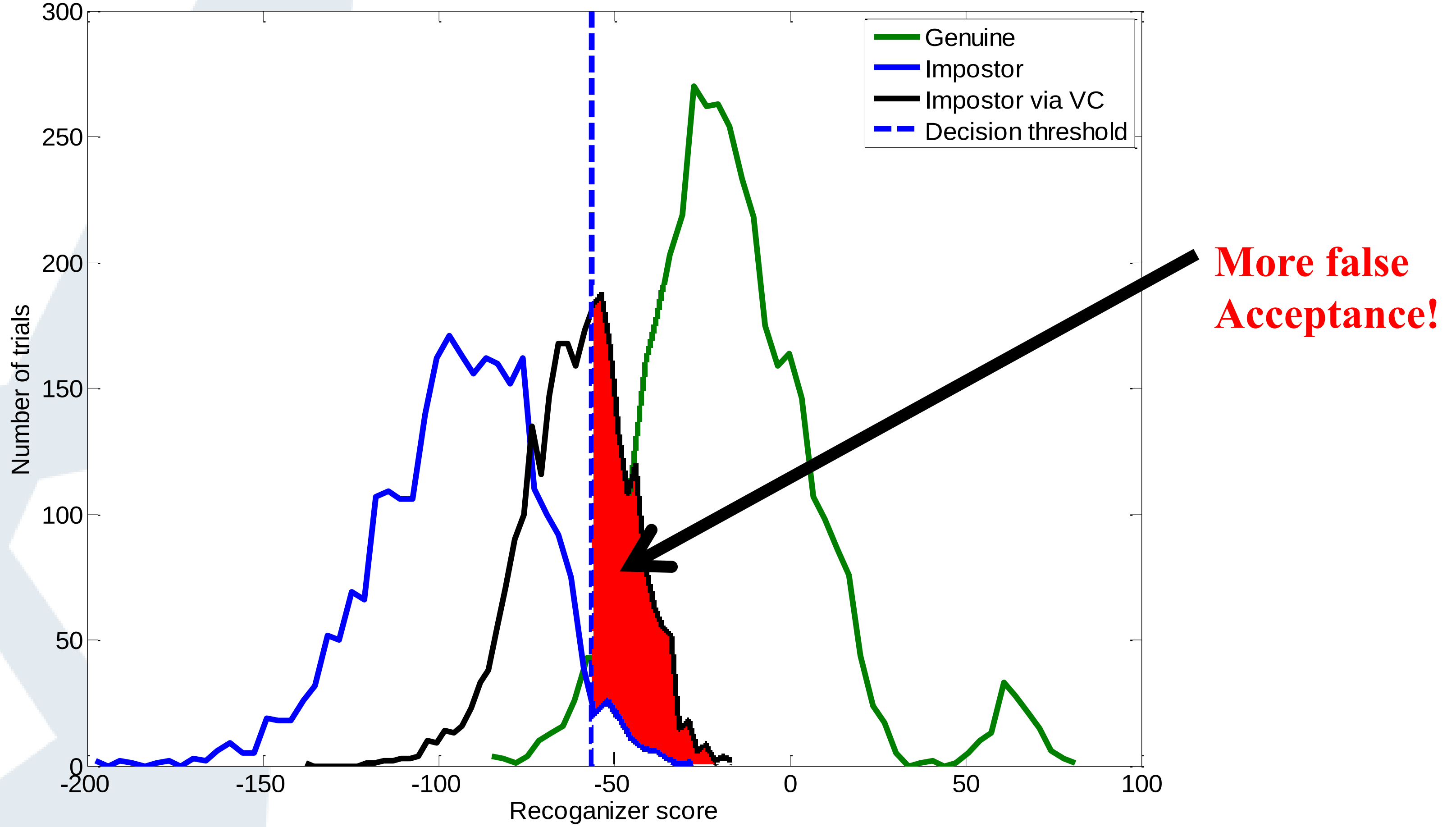
- Dataset design (use a subset of NIST SRE 2006 core task)
- An extreme dataset in which all impostors are voice-converted

	Standard speaker verification	Spoofing attack
Unique speakers	504	504
Genuine trials	3,978	3,978
Impostor trials	2,782	0
Impostor trials (via VC)	0	2,782

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.

Voice Conversion Spoofing Attack

- Score distributions before and after spoofing attack



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.

A summary of spoofing attack studies
(mostly Text-independent test)

Study	VC method	Database	TI or TC or TD	Recognizer	Baseline EER (%)	Spoofing EER (%)	Spoofing FAR (%)
(Bonastre et al., 2007)	FW	NIST SRE 2005	TI	GMM-UBM	8.54	35.41	N. A.
(Bonastre et al., 2007)	FW	NIST SRE 2006	TI	GMM-UBM	6.61	28.07	N. A.
(Alegre et al., 2012a)	FW	NIST SRE 2005	TI	GMM-UBM	8.50	32.60	N. A.
(Alegre et al., 2012a)	FW	NIST SRE 2005	TI	JFA	4.80	24.80	N. A.
(Kinnunen et al., 2012)	JD-GMM	NIST SRE 2006	TI	GMM-UBM	7.63	24.99	N. A.
(Kinnunen et al., 2012)	JD-GMM	NIST SRE 2006	TI	VQ-UBM	7.56	22.62	N. A.
(Kinnunen et al., 2012)	JD-GMM	NIST SRE 2006	TI	GMM-SVM	3.74	12.58	41.54
(Kinnunen et al., 2012)	JD-GMM	NIST SRE 2006	TI	JFA	3.24	7.61	17.33
(Wu et al., 2012c)	US	NIST SRE 2006	TI	JFA	3.24	11.58	32.54
(Wu et al., 2012c)	JD-GMM	NIST SRE 2006	TI	PLDA	2.99	6.77	19.29
(Wu et al., 2012c)	US	NIST SRE 2006	TI	PLDA	2.99	11.18	41.25
(Kons and Aronowitz, 2013)	FW	WF corpus (Aronowitz et al., 2011)	TI	I-vector	1.60	8.80	29.00
(Kons and Aronowitz, 2013)	FW	WF corpus (Aronowitz et al., 2011)	TI	GMM-NAP	1.10	3.40	38.00
(Kons and Aronowitz, 2013)	FW	WF corpus (Aronowitz et al., 2011)	TD	HMM-NAP	1.00	2.90	36.00
(Wu et al., 2013b)	JD-GMM	RSR2015 (Larcher et al., 2012)	TI	GMM-UBM	15.32	25.87	39.22
(Wu et al., 2013b)	US	RSR2015 (Larcher et al., 2012)	TI	GMM-UBM	15.32	27.30	42.56

EER and FAR increase considerably under spoofing attack!

Anthony Larcher and Haizhou Li, The RSR2015 Speech Corpus, IEEE SLTC Newsletter, May 2012

A*STAR I²R

- EER and FAR increase as the number of training utterances for voice conversion increases
- Text-dependent test on RSR 2015 database

	Male		**Female**	
# of training utterances for VC	EER	FAR	EER	FAR
Baseline	2.92	2.92	2.39	2.39
VC 2 utterances	3.90	4.80	1.78	1.06
VC 5 utterances	5.07	9.17	2.51	2.64
VC 10 utterances	7.04	16.20	2.82	3.77
VC 20 utterances	8.30	21.87	3.12	4.68

- Introduction
- Speaker verification
- Voice conversion and spoofing attack
- Anti-spoofing attack
- Future research

Anti-spoofing attack

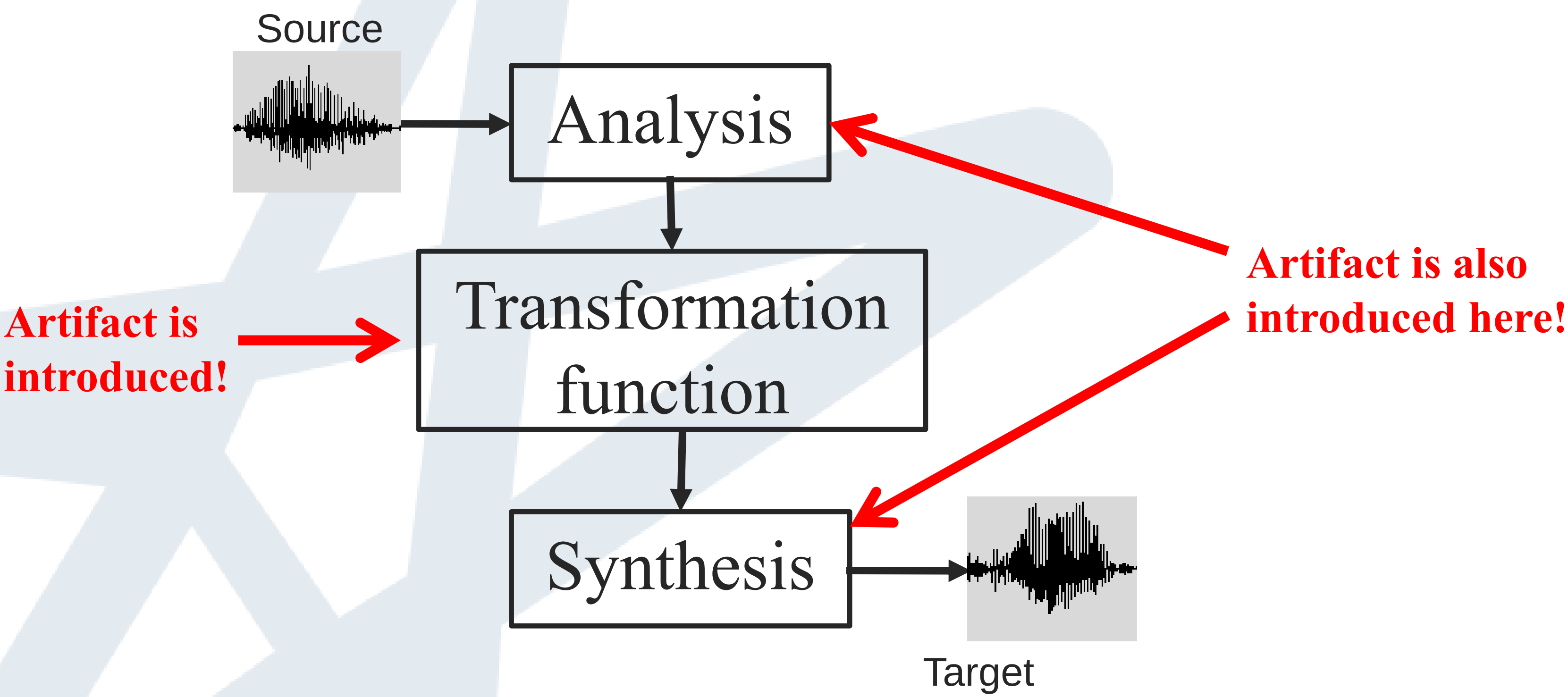
- More accurate speaker verification system is never good enough
  - JFA, PDLA, i-vector
- Synthetic speech detection
  - the absence of natural speech phase [1]
  - the use of F0 statistics to detect spoofing attacks [3]
  - synthetic speech generated according to the specific algorithm [2] provokes lower variation in frame-level log-likelihood values than natural speech
- Countermeasures are specific to a type of synthetic speech, therefore, easily overcome by other voice conversion techniques

1) *Z. Wu, T. Kinnunen, E. S. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific. IEEE, 2012, pp. 1-5*
2) *T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in Proc. Eurospeech, 2001.*
3) *A. Ogihara, H. Unno, and A. Shiozakai, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," IEICE transactions on fundamentals of electronics, communications and computer sciences, vol. 88, no. 1, pp. 280-286, jan 2005*

A*STAR I²R

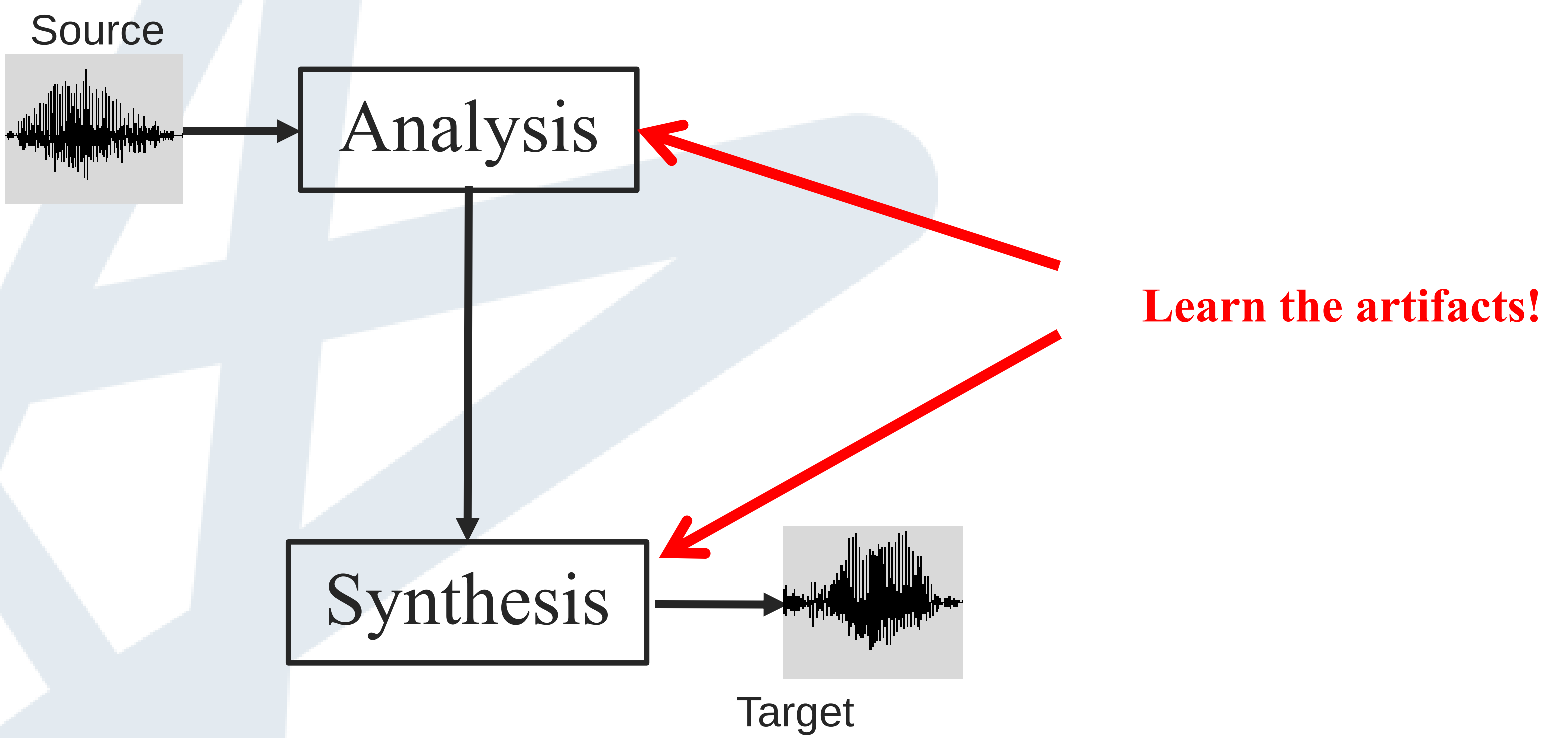
Anti-spoofing attack

- Artifacts are introduced during analysis-synthesis process



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

- Artifacts are introduced during analysis-synthesis process



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

- Natural speech vs copy-synthesis speech

	#1	#2	#3	#4	#5
Natural					
Synthetic					

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

$$X(\omega) = |X(\omega)|e^{j\varphi(\omega)}$$

where $|X(\omega)|$ is the magnitude spectrum and $\varphi(\omega)$ is the phase spectrum.

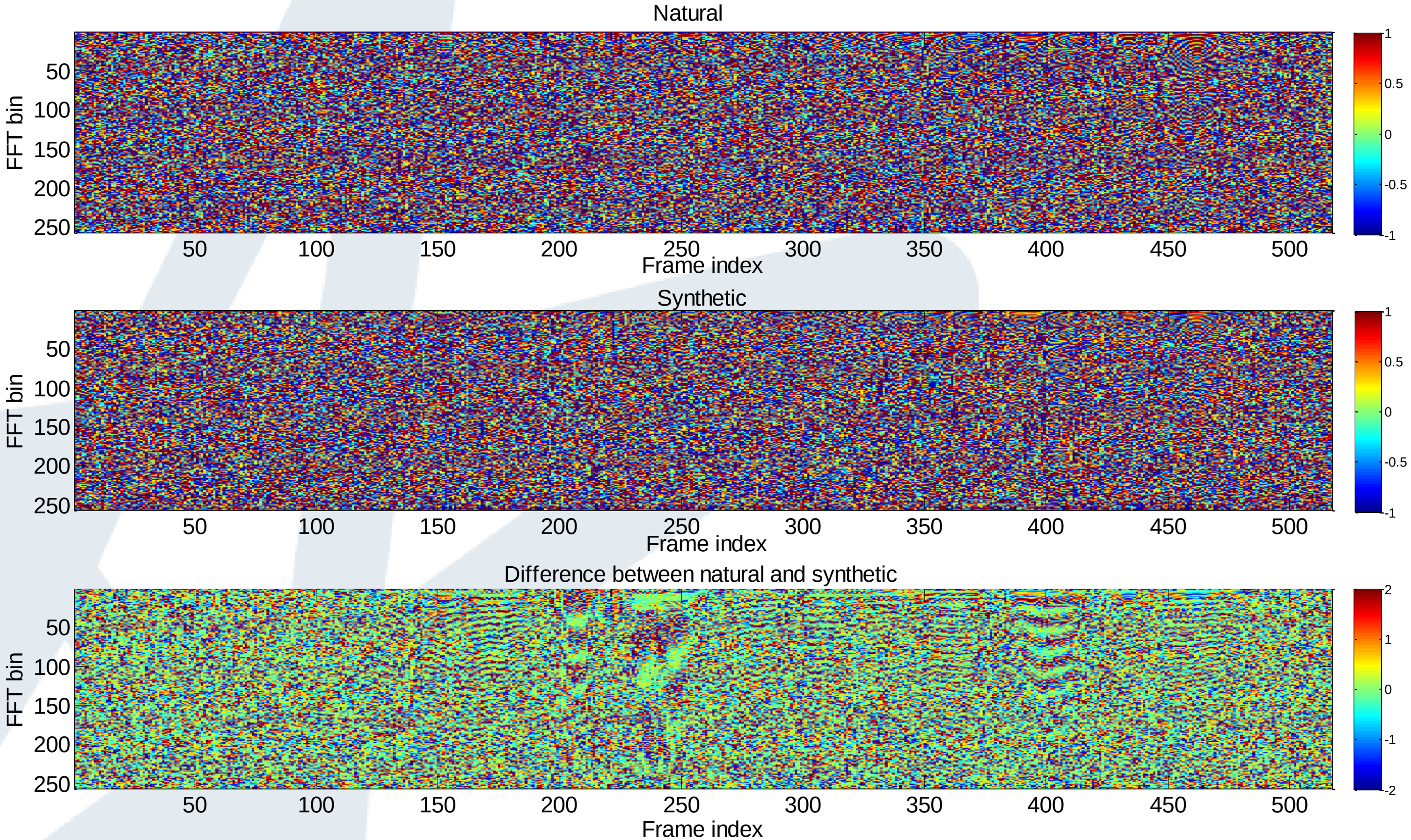
- Cosine-phase spectrum: $\cos(\varphi(\omega))$
- Modified group delay spectrum $\tau_{\rho,\gamma}(\omega)$

$$\tau_\rho(\omega) = \frac{X_R(\omega)Y_R(\omega)+X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\rho}} \qquad \tau_{\rho,\gamma}(\omega) = \frac{\tau_\rho(\omega)}{|\tau_\rho(\omega)|}\tau_\rho(\omega)^\gamma$$

where $X_R(\omega)$ and $X_I(\omega)$ are the real and imaginary parts of $X(\omega)$, respective.
$Y_R(\omega)$ and $Y_I(\omega)$ are the real and imaginary parts of the Fourier transform spectrum of $nx(n)$.
$|S(\omega)|^2$ is the cepstrally smoothed power spectrum.

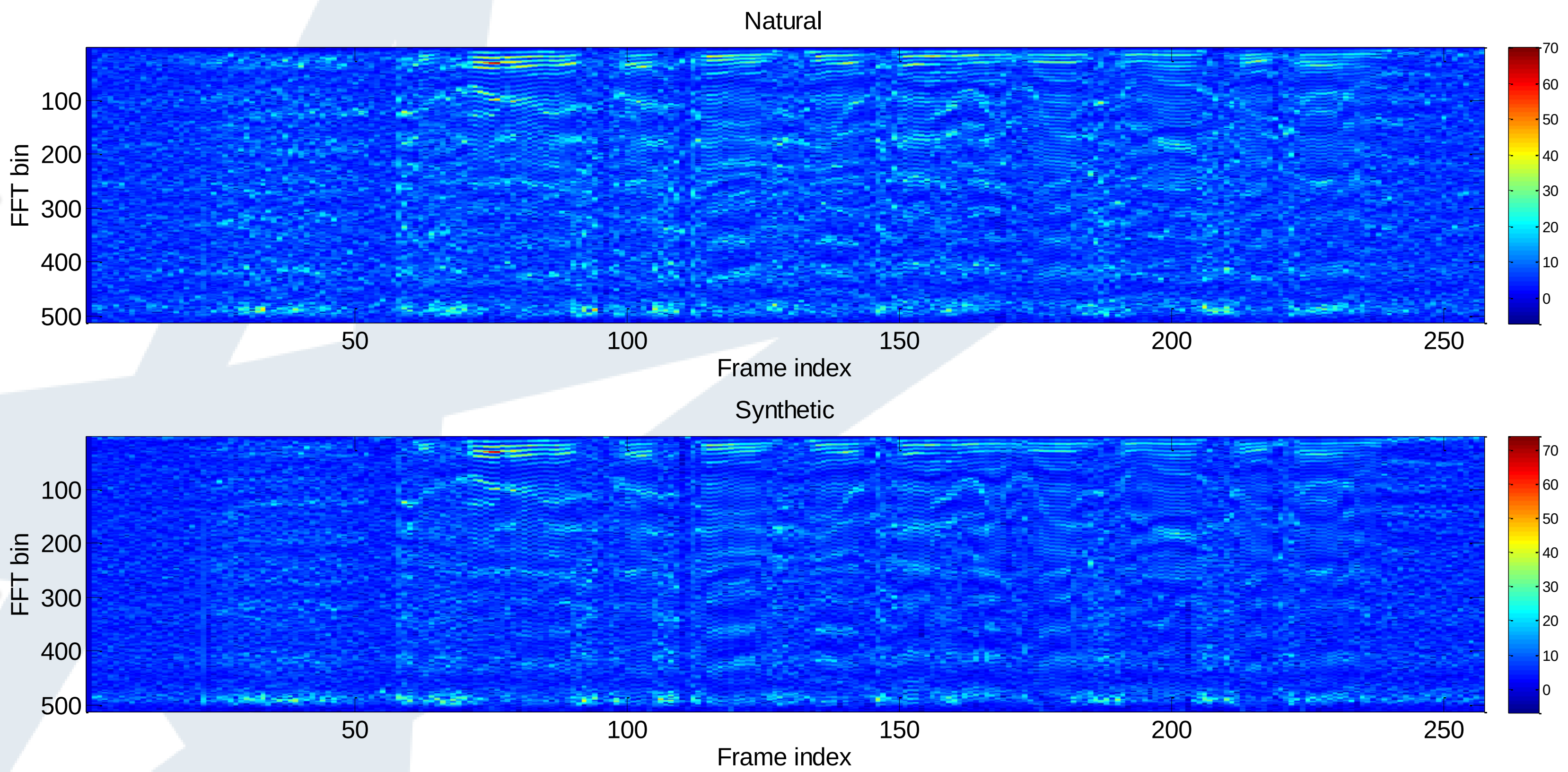
1. Murthy, Hema A., and Venkata Gadde. "The modified group delay function and its application to phoneme recognition." *ICASSP 2003*
2. Hegde, Rajesh M., Hema A. Murthy, and Venkata Ramana Rao Gadde. "Significance of the modified group delay feature in speech recognition." *IEEE Transactions on Audio, Speech, and Language Processing,* 15.1 (2007): 190-202.

- Phase artifacts – cosine-phase spectrogram



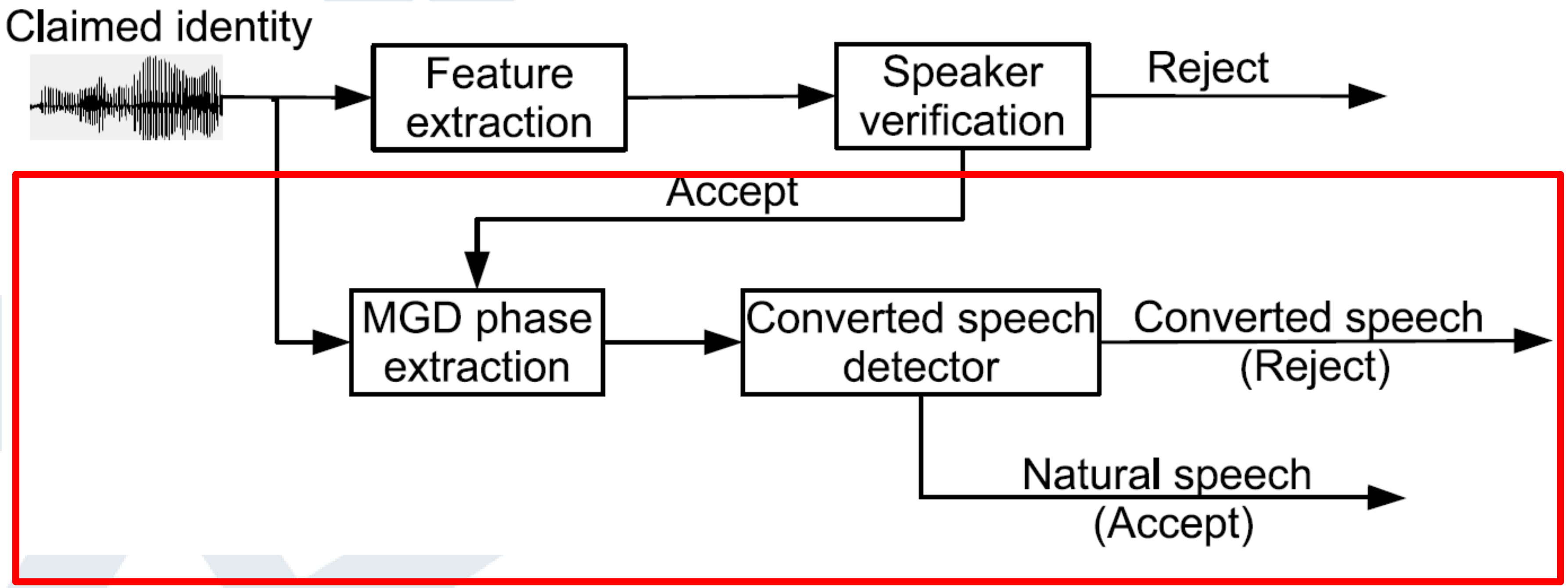
Zhizheng Wu, Eng Siong Chng, Haizhou Li, "Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition", Interspeech 2012

- Phase artifacts – modified group delay spectrogram



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

- Speaker verification system with anti-spoofing countermeasure



Zhizheng Wu, Tomi Kinnunen, Eng Siong Chng, Haizhou Li, Eliathamby Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case", APSIPA ASC 2012.

- Anti-spoofing attack performance

SV system	Voice conversion	False acceptance rate (%)	
		Without anti-spoofing	With anti-spoofing
GMM-JFA	GMM	17.36	0.0
	Unit-selection	32.54	1.64
PLDA	GMM	19.29	0.0
	Unit-selection	41.25	1.71

Zhizheng Wu, Tomi Kinnunen, Eng Siong Chng, Haizhou Li, Eliathamby Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case", APSIPA ASC 2012.

I2R A*STAR

- Introduction
- Speaker verification
- Voice conversion
- Voice conversion spoofing attack
- Anti-spoofing attack
- Future research

Get started!

- Public available resource for spoofing attack studies
  - Voice conversion:
    - Speech signal processing toolkit (SPTK) : http://sp-tk.sourceforge.net/
    - Festvox: http://www.festvox.org/
    - UPC_HSM_VC: http://aholab.ehu.es/users/derro/software.html
  - Speaker verification
    - ALIZE: http://mistral.univ-avignon.fr/index_en.html
  - Datasets for spoofing and anti-spoofing are available upon request
    - http://www3.ntu.edu.sg/home/wuzz/
      - NIST SRE 2006 subset with converted speech
      - WSJ0+WSJ1 for anti-spoofing
  - A special session was organized in INTERSPEECH 2013 conference on Spoofing and Countermeasures for Automatic Speaker Verification