Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification

Sayaka Shiota（Tokyo Metropolitan University）, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono,
Isao Echizen（National Institute of Informatics）, Tomoko Matsui（Institute of Statistical and Mathematics）

1. Introduction

Automatic speaker verification (ASV)

- An easy-to-use biometric authentication system
- State-of-the-art system: i-vector, PLDA
  - Show potential to support mass-market adoption

Speech Synthesis Techniques (Text-to-speech; TTS)

- Generate natural-sounding artificial speech with targeted speaker's few voices.
- State-of-the-art system: HMM-based, Voice conversion
  - Help individuals with vocal or communicative disabilities

Spoofing attacks against ASV system

- ASV performance is seriously degraded.
- Main types of spoofing attacks:
  Replay, Speech synthesis, Voice conversion
- Some anti-spoofing techniques have been reported.

A fundamental solution against the spoofing attacks is required.

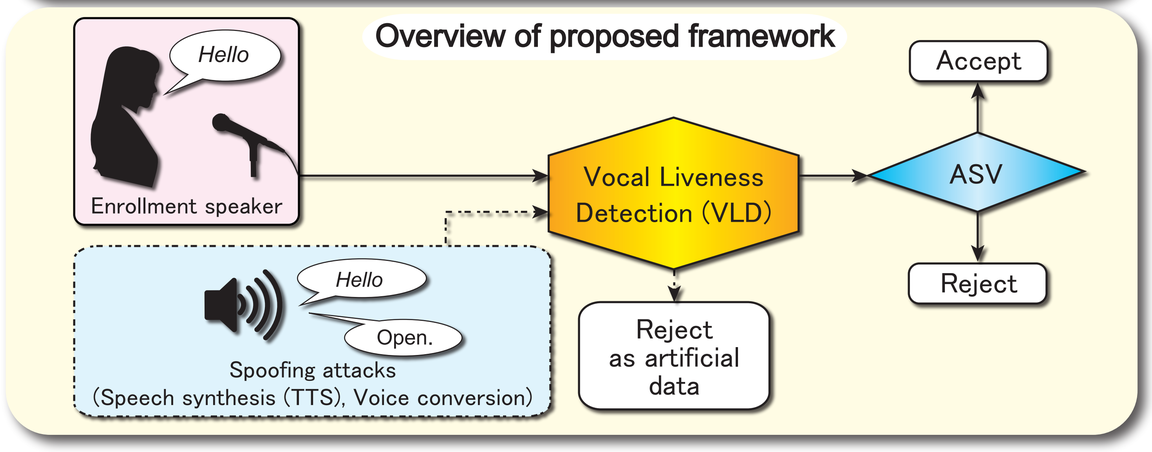
2. Voice Liveness Detection (VLD)

Procedures for spoofing attacks

- Play spoofing speech via loudspeakers

Distinguish input data produced by a live human from input data played via loudspeakers.

Can protect against all types of spoofing attacks



What is the liveness evidence in a speech waveform?

- Voice made by airflow, and it transform to an acoustical signal via a microphone
- **Pop noise phenomena**: a sort of perceived plosive burst
  Only living human caused pop noise.

Pop noise detection leads to reduce the vulnerability of ASV

3. Pop noise detection algorithms

Low-frequency-based single channel detection

- Pop noise appears as high energy regions at very low freqency (Fig.1)
  - Sudden irregular modulations of strong energy
  - Durations typically rangin between 20 - 100 msec
- A min/max energy variation and velocity ensure there will be a relative increment/drop in the pop noise energy.

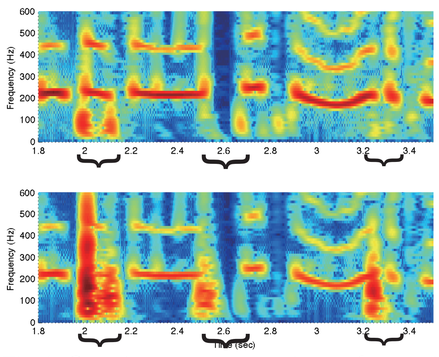


Fig.1, Spectrogram comparison of recording using (top) or not using pop filter (bottom).

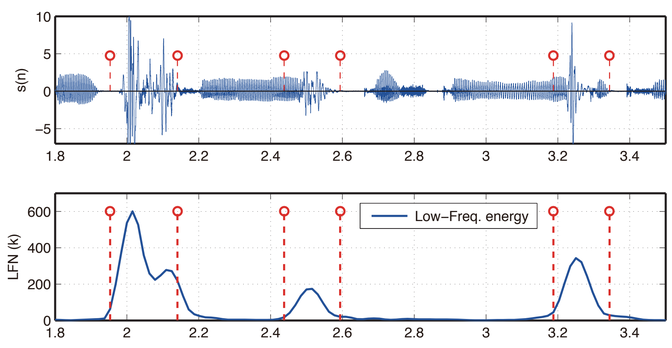


Fig.2, Example of pop noise detection. Time-domain signals(top), average low-band energy(bottom), and the detected pop noise boundaries (red dotted).

Subtraction-based detection with two channels

- Capture the whole freq. components of the pop noise
- Two microphone are used.
  - only one of them has a pop filter (Fig.3)

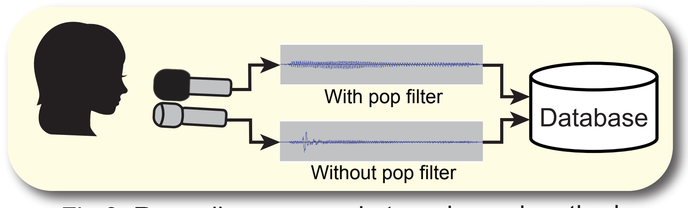


Fig.3, Recording process in two channel method

- Assuming only one signal includes pop noise, it is estimated by subtracting the ordinary speech component as follows:

$$D(b,\omega) = F_p(b,\omega) - C(\omega)F_x(b,\omega)$$

Non-Filtered speech / Filtered speech

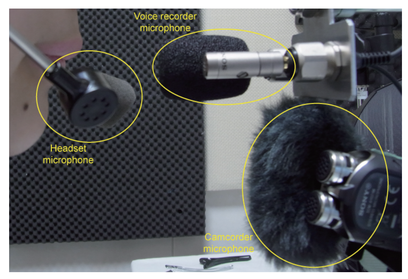
$F(b,\omega)$: STFT
b : Time frame
ω : Angular frequency

- $C(\omega)$ represents a compensation filter between freq. characteristics of the two channels.
- An estimate of $C(\omega)$ to minimize $\sum_{b,\omega}|D(b,\omega)|^2$ can be represented as $C(\omega) = \frac{\sum_b F_p(b,\omega)F_x(b,\omega)^*}{\sum_b |F_x(b,\omega)|^2}$. (* complex conjugate)
- Amplitude of inverse STFT for $D(b,\omega)$ is used for estimating boundaries of the pop noise.

5. Experiments

Experimental conditions

- Database including pop noise is recorded with three kinds of microphones
  - Compatible microphone with camcorder (CAMCORDER)
  - Microphone with a voice recorder (VOICE)
  - Microphone with a headset (HEADSET)
- 17 female Japanese speaker
- 100 sentences for each speaker
- 48kHz sampling
- Training data: 70 sentences per speaker
- Test data: 30 sentences per speaker
- Spoofing data: 31 sentences estimated by HTS adaptation technique



Experimental results

1.Pop noise detection test

- Judge an input signal comes from a live human or a loudspeaker.
- Both method can capture pop noise as liveness evidence
- Pop noise phenomenon depends on the microphone type

		Input	
		Human	Spoof
Judge	Human	Correct	FAR
	Spoof	FRR	Correct

Tab. 1, EER (FAR=FRR) of VLD algorithms with some microphone

Microphone	Single ch.	Two ch.
VOICE	4.73%	29.11%
CAMCORDER	36.06%	45.52%
HEADSET	3.95%	5.88%

2.Combine VLD module and ASV system (VLD+ASV)

- Judge an input signal comes from a live human or a loudspeaker, and judge the input signal is a enrollment speaker or not.

Tab. 2, EER of the ASV system
w/ SA: test data includes spoofing attack data, w/o SA: test data includes no spoofing attack data

Microphone	w/o SA	w/ SA	VLD+ASV	
			Single ch.	Two ch.
VOICE	5.49%	5.53%	5.48%	5.49%
CAMCORDER	4.69%	6.61%	5.23%	5.30%
HEADSET	4.28%	6.61%	4.45%	4.28%

- ASV performance is degraded by SA data
- VLD+ASV performance is almost same as no SA system

Pop noise detection algorithm works well as VLD module

6. Conclusion

- VLD algorithms can reduce the vulnerabilities of ASV against to spoofing attacks.
- Future work: Use larger database, Evaluate other spoofing attacks (e.g., VC, Unit selection), Distinguish pop noise from wind noise.