VOICE LIVENESS DETECTION FOR SPEAKER VERIFICATION BASED ON A TANDEM SINGLE/DOUBLE-CHANNEL POP NOISE DETECTOR

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

Spoofing attacks against ASV systems

- 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.

Voice liveness detection (VLD) framework [Shiota; '15]

- Identification of liveness information in input samples
- High performance for detecting spoofinging attacks

Pop noise detector

- Capture pop noise as liveness evidence
- Single- and double-channel approaches were reported

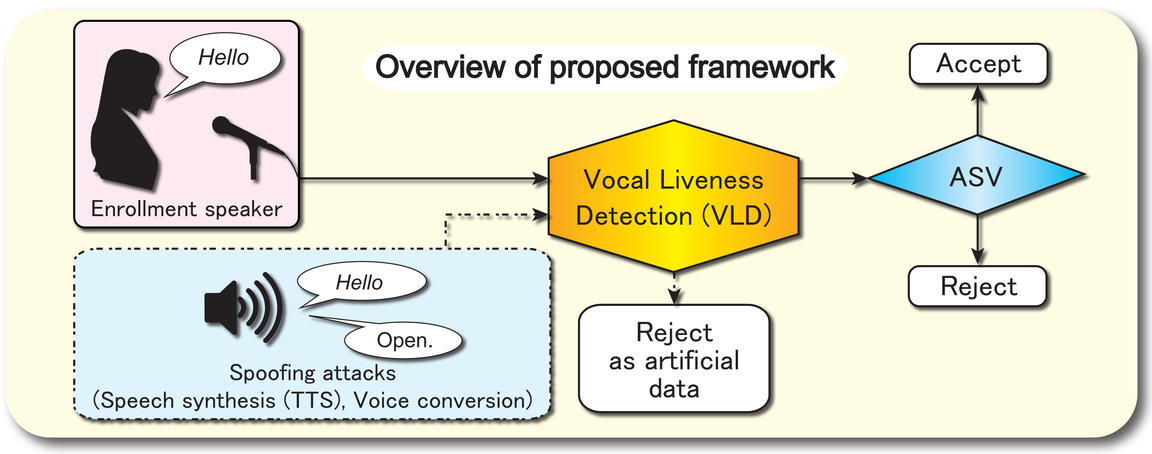
Tandem approach leads to improve the VLD performance

2. Voice Liveness Detection (VLD)

Concept of VLD framework

- Spoofing attacks are played via loudspeakers
- Distinguish input data produced by a live human from input data played via loudspeakers.

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

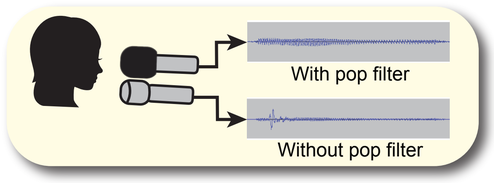
3. Pop noise detection algorithms

Low-frequency-based single channel algorithm

- Pop noise appears as high energy regions at low freqency
  - Sudden irregular modulations of strong energy
  - focus under 1000Hz
- A min/max energy variation and velocity ensure there will be a relative increment/drop in the pop noise energy.

Subtraction-based detection with double channel

- Two microphone are used.
- Capture the whole freqency components of the pop noise



- 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

- 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.

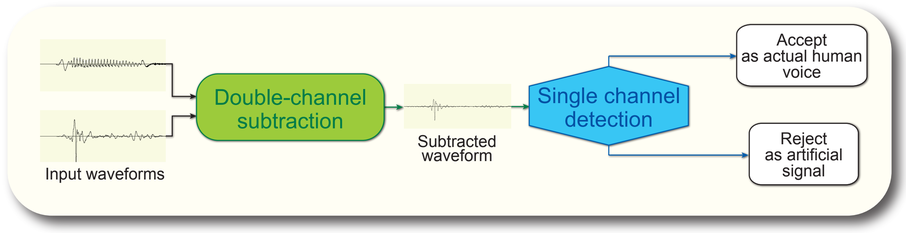
4.Tandem single/double-channel popnoise detection algorithm

Benefits and drawbacks of the conventional algorithms

- Single-channel algorithm
  - Performance was depended on speakers and microphones
  - Difficult to distinguish played pop noise from real pop noise
- Double-channel algorithm
  - Amplitude-base decision can characterize presence of the pop noise more precisely
  - Performance is lower than the single algorithm

Tandem algorithm

- First step: double channel subtraction used to obtain subtracted waveform
- Second step: single channel detection

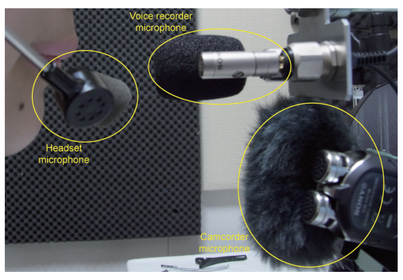


Subtracted waveform has irregular modulations precisely and it help to improve the single-channel detection

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: 30 sentences estimated by HTS adaptation technique



Experimental results

1.Pop noise detection performance

- Judge an input signal comes from a actual human or a loudspeaker.

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

		Input	
		Human	Spoof
Judge	Human	Correct	FAR
	Spoof	FRR	Correct

- Tandem method can capture pop noise as liveness evidence
- Pop noise phenomenon depends on the microphone type

Tab. 2, EER of the VLD algorithms

Microphone	Single	Double	Tandem
VOICE	36.06%	45.52%	26.61%
CAMCORDER	4.73%	29.11%	0.95%
HEADSET	3.95%	5.88%	2.35%

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.

Fig 1, EER of the ASV + VLD

- ASV performance is degraded by spoofing attacks
- Tandem approach obtains high performance

6. Future work

- Conducting trials using a larer database
- Simple pop noise detection can be broken by wind or spoof breath. Thus, considering phoneme information in pop noise periods.
- Evaluate the replayed pop noise samples via each pop noise detection algorithms.