[ICASSP 2020]

Detecting Mismatch between Text Script and Voice-over Using Utterance Verification Based on Phoneme Recognition Ranking

Yoonjae Jeong and Hoon-Young Cho
Speech Lab, AI Center, NCSOFT

Contents

Introduction – Motivation

Massive Amounts of Text Scripts and Voice-overs

Popularization of Computer Games → A Large Number of Text Scripts and Their Voice-overs

Captions and Voice-overs of NPCs in an MMORPG, Blade & Soul of NCSOFT

Creation and Verification Process of Voice-overs

Text Scripts → Outsourcing to Voice Actors → Voice-overs → Manual Verification

Text Scripts → Manual Verification

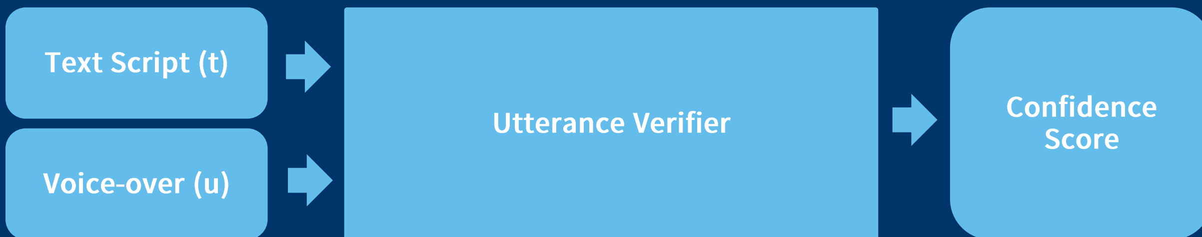
At Least 10Ks of Voice Resources for a Language

Extremely Time Consuming & Labor-Intensive

Introduction – Automation of Verification Process

Utterance Verification (UV) [Jiang, 2005]

- One of the key technologies to automate the verification process.



- The confidence is based on the gap of phoneme recognition probabilities between an acoustic model (H0) and its anti-phoneme model (H1).

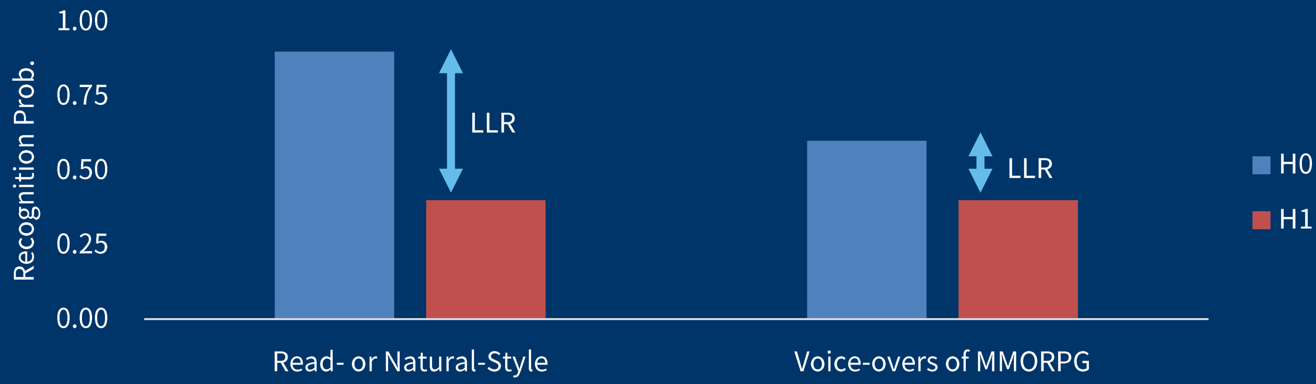
$$LLR(t, u) = log \frac{P(H_0)}{P(H_1)}$$

Introduction – Problem of Conventional UV

Difference of Speech Style

	Speech Style	Example
Acoustic Model	Read- or Natural-style	Excerpted from LibriSpeech [Panayotov et al., 2015]
Voice-overs	Exaggerated and Emotional Intonation	Excerpted from Blade & Soul [NCSOFT]

Decrease of the Probability Gap for Voice-overs



Introduction – Proposed Solution

Average Phoneme Recognition Ranking (APR) Based UV

[Observation]

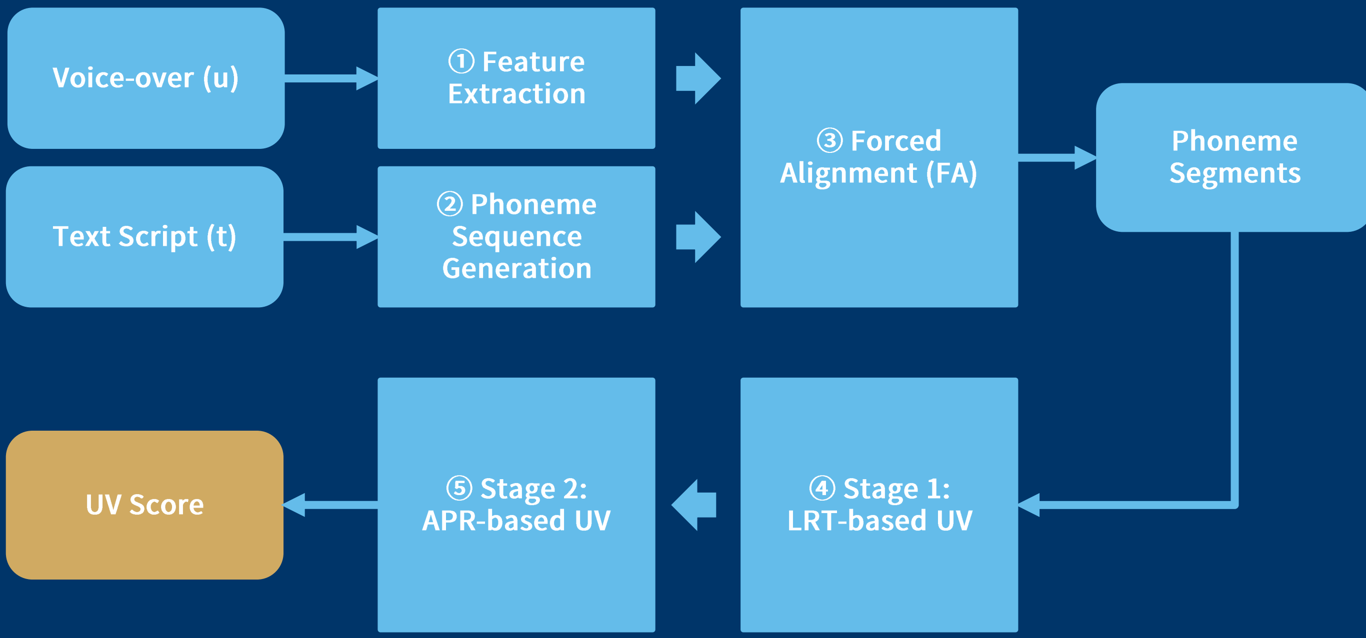
The phoneme recognition rankings do not significantly change regardless of the speech styles.

[Proposed APR-based UV]

The average phoneme recognition ranking of each speech segment of a phoneme sequence corresponding to its text script as the confidence measure

Proposed Method

Procedure of Proposed UV System

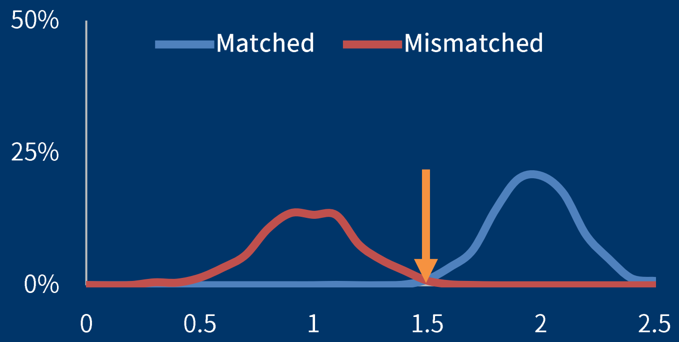


Proposed Method – Basic LRT-based UV

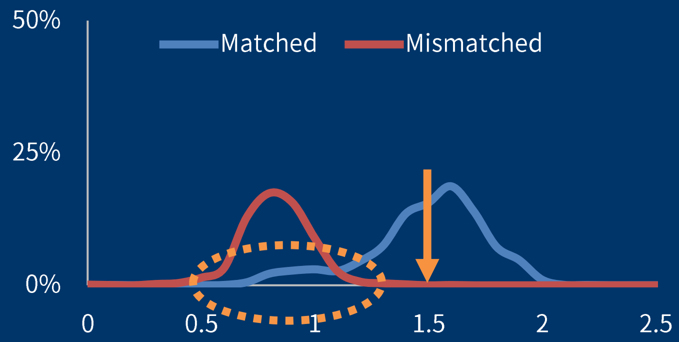
Likelihood Ratio Test (LRT) based Utterance Verification

$$LLR(t,u) = log\frac{P(H_0)}{P(H_1)} = g(t,u) - G(t,u) > \tau$$

Distribution of LLR for the Script-Voiceover Pairs



Read- or Natural Style



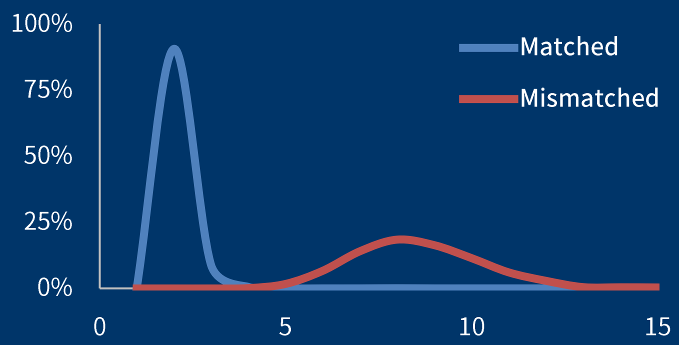
MMORPG

Proposed Method – Proposed APR-based UV

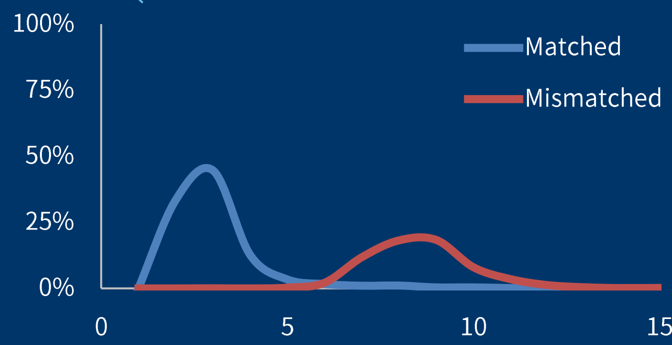
Average Phoneme Ranking (APR) based Utterance Verification

$$APR(t,u) = \frac{1}{N}\sum_{i=1}^{N} rank(p_i, f_i) < \theta$$

Distribution of APR for the Script-Voiceover Pairs



Read- or Natural Style



MMORPG

Proposed Method – Complement for Rare Errors

Two-stage APR-based Utterance Verification

- Some rare cases where the phoneme recognition rankings are high although the overall recognition probabilities are extremely low.

- To avoid the occurrence of such scenarios:

$$APR_{2-stage}(t,u) = \begin{cases} |P| & \text{if } LLR(t,u) \leq \tau \\ APR(t,u) & \text{otherwise} \end{cases} < \theta$$

Experiment – Test Sets

[Test Set 1] WSJ-CAM0 Corpus Test Set

- Comparison with the State-of-the-Art [Huang & Hain, 2019]
- Mismatched Samples Are
  - Randomly deleting (Del), inserting (Ins), and substituting (Sub) four word.

[Test Set 2] For Detecting a Mismatch between Text Script and Voice-over

- Excerpted Test Sets from a Korean Speech Database (DICT01) and an MMORPG (BNS)

Test Set	Description	Matched	Mismatched
DICT01	Read-style	1,600	1,600
BNS-1	Exaggerated-style	1,600	1,600
BNS-2	Exaggerated-style + various tones & effects	483	483

Experiment – Evaluation (1/4)

Comparison with Previous Work

- Comparison of 4-word mismatch detection accuracy for deletion (Del), insertion (Ins), and substitutions (Sub) in the WSJ-CAM0 test set.

	Del	Ins	Sub	Average
Cross-model Attention [Huang & Hain, 2019]	0.781	0.792	0.558	0.710
LRT [Rahim, Lee & Juang, 1997]	0.605	0.798	0.670	0.691
Proposed APR	0.730	0.986	0.918	0.878
Proposed $\text{APR}_{\text{2-stage}}$	0.731	0.986	0.920	0.879

Performance Decrease in the Del errors, but Much More Improvement for the Ins and Sub Errors.

Experiment – Evaluation (2/4)

Performances for Detecting Mismatch between Text Script and Voice-over

- Comparison between the proposed APR-based UV and the conventional LRT-based UV with the optimized thresholds.

Test Set	LRT		APR		
	ACC	τ	ACC	θ	Δ
DICT01	0.992	1.5	0.998	4.0	+0.006 (0.6%)
BNS-1	0.930	1.2	0.968	5.0	+0.038 (4.1%)
BNS-2	0.901	1.1	0.959	6.0	+0.058 (6.4%)

Significant Improvement in Exaggerated Voice-overs

Experiment – Evaluation (3/4)

Robustness to Threshold

- Performance degradation in the exaggerated voiceovers, when applying the optimized thresholds of the read speech utterances.

Test Set	LRT		APR	
	ACC	Δ	ACC	Δ
BNS-1	0.813	-0.117 (-14.4%)	0.952	-0.016 (-1.7%)
BNS-2	0.674	-0.228 (-33.8%)	0.900	-0.059 (-6.6%)

Remarkably Lower Performance Drops for the Exaggerated Voice-overs

Experiment – Evaluation (4/4)

Effects of Two-stage Approach

- Performance improvement of the two-stage APR-based UV.

Test Set	APR	$APR_{2\text{-stage}}$	Δ
BNS-1	0.9675	0.9677	+0.0002
BNS-2	0.9592	0.9598	+0.0006

Compensation for a Few Errors of the Pure APR-based UV

Conclusions and Future Work

Conclusions

- We proposed a novel APR-based UV method.
- Performance improvements are over the state-of-the-art.
- Only a small amount of performance degradation with exaggerated voiceovers, even though the model is optimized to read-style utterances.

Future Work

- Handling of Deletion Errors
  - The proposed APR-based UV showed performance degradation for missing words when compared to the state-of-the-art.
- Handling of Laughing-style Utterances
  - Since laughing-style utterances are pronounced differently depending on the situation.
  - Transcribing them to proper phoneme sequences is a challenging task.

References

- T. Hain and O. Saz, “Factored WSJCAM0 Speech Corpus,” [Online]. Available: https://mini.dcs.shef.ac.uk/resources/wsjcam0/, 2013.

- Q. Huang and T. Hain, “Detecting Mismatch Between Speech and Transcription Using Cross-Modal Attention,” in *Proceedings of Interspeech 2019*, 2019, pp. 584–588.

- H. Jiang, “Confidence measures for speech recognition: A survey,” *Speech Communication*, vol. 45, no. 4, pp. 455–470, 2005.

- NCSOFT, “Blade & Soul,” [Online]. Available: http://bns.plaync.com/.

- V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *Proceedings f the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

- M. G. Rahim, C.-H. Lee, and B.-H. Juang, “Discriminative utterance verification for connected digits recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 266–277, 1997.

Q & A

Contacts

Yoonjae Jeong:	yjeong@ncsoft.com
Hoon-Young Cho:	hycho@ncsoft.com