PH2-11 音声認識出力の曖昧性に頑健な音声翻訳のための音声認識の精度ごとの性能比較

胡 尤佳[1], 須藤 克仁[1,2], Sakriani Sakti[1,2,3], 中村 哲[1,2]

[1]奈良先端科学技術大学院大学 [2]理化学研究所革新知能統合研究センターAIP [3]北陸先端科学技術大学院大学

概要

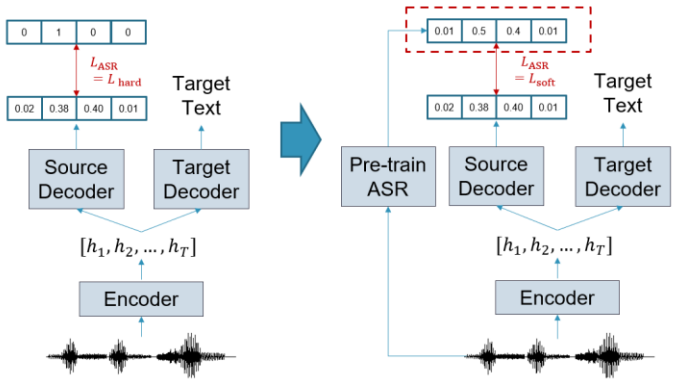
▷**手法** 音声認識(ASR)出力をreferenceとした音声翻訳(ST)
→ 人間で起きうる発音による聞き間違えやすさの考慮が可能

▷**本研究での分析**

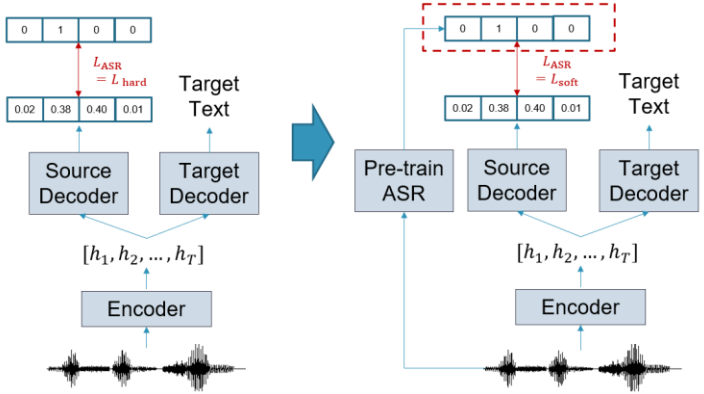
- 音声認識モデルの精度による音声翻訳への影響
- ST-task出力とASR-task出力の比較

関連手法

▷**どれくらい他単語と聞き間違えやすいかを考慮した音声翻訳**
ref: ASR出力確率分布のsoftmax / L_{ASR} : ASR-PBL



▷**どの単語と聞き間違えやすいかを考慮した音声翻訳**
ref: ASR出力のone-best / L_{ASR} : ASR-SBL

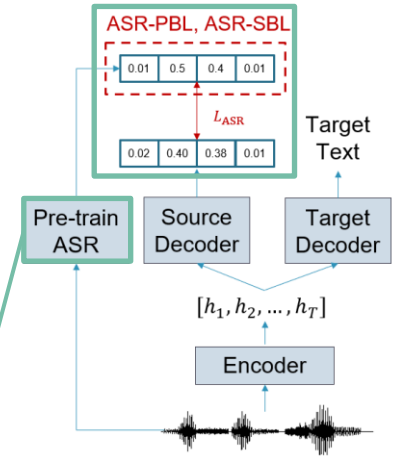


実験

▷**データ** : Fisher Spanish (Es-En), MuST-C (En-De) (論文参照)
▷**実装**: ESPnet, Transformer
▷**ASR-task loss**
▷**Baseline : Cross entropy loss (CE), CE+Label smoothing (CE-LSM)**
▷**ASR Posterior-based loss (ASR-PBL)** $\lambda_{soft} = \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$
▷**ASR Sequence-based loss (ASR-SBL)** $\lambda_{soft} = \{0.25, 0.5, 0.75, 1.0\}$ (w/, w/o LSM)
$\lambda_{ASR} = \{0.5\}$ $L_{ASR} = \lambda_{soft} L_{soft} + (1 - \lambda_{soft}) L_{hard}$ $L = \lambda_{ASR} L_{ASR} + (1 - \lambda_{ASR}) L_{ST}$

▷**分析**
▷**ASR-SBLの有効性 (ASR-PBLと同様に効果的か)**
▷**異なるWERのPre-train ASRで比較**
▷**ASR-taskの出力の比較**

	ASR model	Soft label WER
High WER	Epoch-6	39.7
	Epoch-8	35.3
	Epoch-10	28.9
	Attn-Specaug	14.1
	Attn	9.3
Low WER	Attn-CTC	7.8
	Attn-CTC-Specaug	6.7



結果

	WER	BLEU			
		ASR-PBL		ASR-SBL (LSM)	
ASR model	Soft-label	soft-0.5	soft-1.0	soft-0.5	soft-1.0
None (Single-task)	-	40.66			
None (CE)	-	43.83			
None (CE-LSM)	-	45.16			
Epoch-6	39.7	**45.34**	41.71	44.68	41.71
Epoch-8	35.3	**45.33**	43.54	**45.63**	42.34
Epoch-10	28.9	**45.76**	42.93	**45.86**	43.79
Attn-Specaug	14.1	**45.29**	43.73	**45.34**	44.34
Attn	9.3	**46.04**	44.53	**45.35**	44.03
Attn-CTC	7.8	44.93	43.98	44.87	44.40
Attn-CTC-Specaug	6.7	44.66	44.82	**45.75**	44.76

表1 : soft-{0.5, 1.0}でのBLEU

- ASR-SBLもASR-PBLと同様にsoftとhardを同じほどの割合で学習させたときに効果的
- High WERのものもsoft-0.5で効果的に利用できる
- 直接WERの影響を受けるのはsoft-1.0のみ

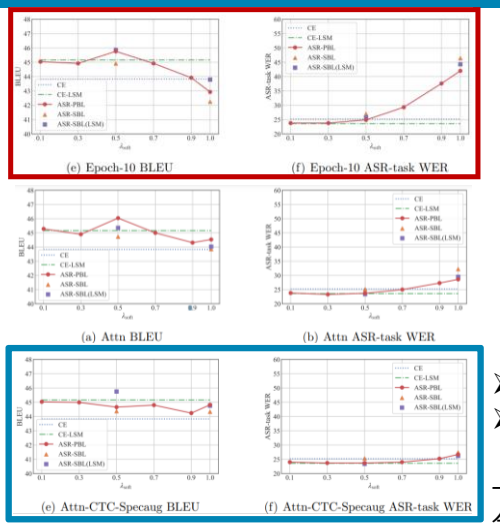


図1: 異なるPre-train ASRでのそれぞれのλ_{soft}でのBLEUとASR-task WER

- ほとんど場合でsoft-0.5が効果的

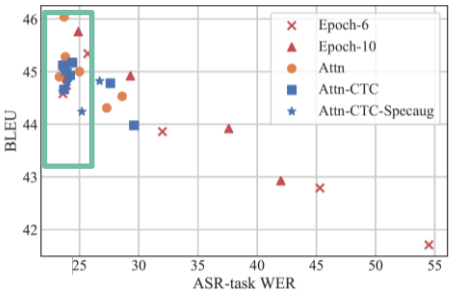


図2: ASR-PBLでのBLEUとASR-task WER (λ_{soft}は考慮しない)

- 傾向: ASR-task WER低→BLEU高
- Low WER best BLEU より High WER best BLEUの方がよい

→ Lowがsharp, Highがsmoothな分布で，smooothな分布が効果的?

▷**今後の方針**
soft-label decodingの負荷の削減
Pre-train ASRの学習中の精度の調整