Multi-task Learning is Not Enough: Representational Entanglement in Dual-output Second Language Speech Recognition

Seung Hwan Cho [1] **Young-Min Kim** [1 2]

Abstract

Second-language (L2) speech recognition often requires transcriptions of pronunciations and intended meanings. Multi-task learning (MTL) is a natural approach because it assumes that shared representations benefit both outputs. However, this paper shows that this assumption does not hold across Korean and English. MTL improves meaning but degrades surface transcription, especially in English, where the degradation scales with surface-meaning divergence measured by Levenshtein edit distance. Encoder analysis links these patterns to encoder-level entanglement, with Korean preserving disentangled representations while English produces nearly identical ones. Cross-output decoder analysis shows that the meaning dual-output decoder adapts with a unique representation, while the surface dual-output decoder remains constrained by the encoder. These findings motivate the design of MTL frameworks that mitigate encoder-level entanglement to reduce surface degradation in dual-output L2 automatic speech recognition.

1. Introduction and Related Work

Human speech exhibits systematic differences between what is actually pronounced (surface-level) and the canonical written form (meaning-oriented) of an utterance. These differences reflect language-specific phonological phenomena, such as coarticulation, phonological reduction, and liaison (Ernestus & Warner, 2011). The phonological gap is more pronounced in second-language (L2) speech, where speaker-specific deviations are common (Munro, 2021). Therefore, speech recognition systems for L2 speakers must recover both transcription forms from a single acoustic signal to enable targeted feedback in language learning and pronunciation assessment applications (Eskenazi, 2009).

Multi-task learning (MTL) is a natural approach for dual-output (DO), encompassing auxiliary MTL where one task supports another and joint MTL where tasks are learned with equal importance (Ruder, 2017). In automatic speech recognition (ASR), joint connectionist temporal classification (CTC)-attention training (Kim et al., 2017; Watanabe et al., 2017) and intermediate-layer CTC (Nozaki & Komatsu, 2021) use auxiliary CTC to improve alignment and training stability. Joint MTL approaches generate distinct target sequences from the same acoustic input. Examples include dual-decoder models for ASR and speech translation (Le et al., 2020) and unified diarization-separation-recognition systems (Shakeel et al., 2025). However, the effectiveness of joint MTL for DO L2 ASR remains unexamined, despite the fact that the two outputs share linguistic content.

This paper challenges the assumption that joint MTL benefits both outputs in DO ASR by conducting controlled experiments on Korean and English L2 speech. The results show that joint MTL produces asymmetric output trade-offs that depend on language, with the source localized to encoder-level representational entanglement. Our contribution is twofold. First, we demonstrate the language-dependent behavior of joint MTL for DO L2 ASR. Second, we identify the underlying mechanisms through encoder and cross-output decoder analyses. These findings motivate the development of structured approaches to mitigate this entanglement.

2. Method

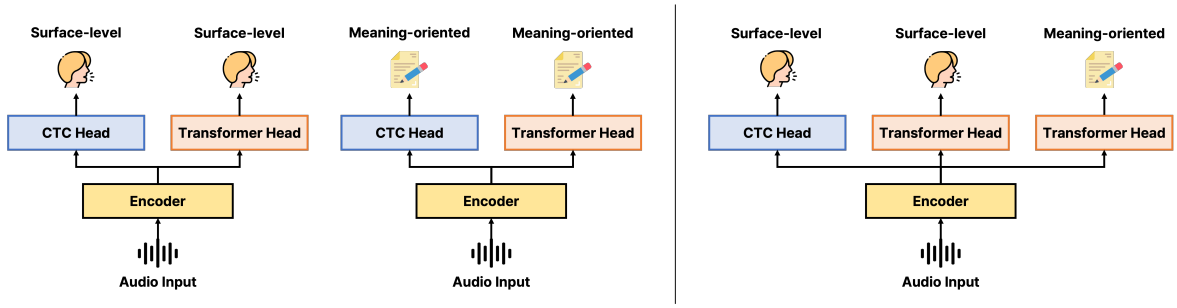
To isolate the effect of joint training on shared representations, we compare single-output (SO) models, which are trained separately on each output, and DO models, which are trained jointly. This comparison is illustrated in Figure 1.

2.1. Problem Formulation

Given an input log mel-spectrogram $X \in \mathbb{R}^{T\times 80}$ with T frames, two output sequences are produced. The surface-level transcription y^{surf} represents the verbatim spoken form,

[1]Department of Industrial Data Engineering, Hanyang University, Seoul, South Korea [2]School of Interdisciplinary Industrial Studies, Hanyang University, Seoul, South Korea. Correspondence to: Young-Min Kim <yngmnkim@hanyang.ac.kr>.

Accepted at the 43^{rd} International Conference on Machine Learning Workshop on Machine Learning for Audio, Seoul, South Korea. 2026.



(a) Single-output configuration

(b) Dual-output configuration

Figure 1. (a) Single-output configuration trains separate models for surface-level (left) and meaning-oriented (right) transcription, each with its own encoder, decoder, and auxiliary CTC head. (b) Dual-output configuration shares a single encoder with two separate Transformer decoders that jointly produce both outputs, along with an auxiliary CTC head on the encoder output.

while the meaning-oriented transcription y^{mean} represents the intended written form. Each token is drawn from a fixed vocabulary $\mathcal{V}$.

2.2. Architecture and Training

Single-output Model. Figure 1(a) shows the SO baseline, which follows the hybrid CTC-attention paradigm with an auxiliary CTC head on the encoder output for alignment supervision. Separate models are trained for surface-level and meaning-oriented transcription. The training objective is $\mathcal{L}_{\text{single}} = \alpha\mathcal{L}_{\text{CTC}} + (1-\alpha)\mathcal{L}_{\text{att}}$ with $\alpha = 0.2$, where $\mathcal{L}_{\text{CTC}}$ is the CTC loss on the target transcription, $\mathcal{L}_{\text{att}}$ is the cross-entropy loss of the attention decoder.

Dual-output Model. The DO model uses one encoder and two decoders, one for each transcription type shown in Figure 1(b). Both decoders attend to the same encoder output through audio cross-attention, and the auxiliary CTC head is retained on the encoder output. The auxiliary CTC is trained on surface-level targets, which preserve the monotonic frame-token alignment that CTC assumes. The training objective is $\mathcal{L}_{\text{dual}} = \alpha\mathcal{L}_{\text{CTC}} + \beta\mathcal{L}_{\text{surf}} + \gamma\mathcal{L}_{\text{mean}}$ with $(\alpha, \beta, \gamma) = (0.2, 0.5, 0.3)$, all fixed via pre-experiments on the validation set. By keeping the auxiliary CTC identical in both configurations, the only architectural difference lies in the presence of the second decoder.

3. Experiments

3.1. Experimental Setup

Dataset. The two AI-Hub[1] datasets, "Educational Korean Audio Data Recorded by Native (L1) Chinese and Japanese Speakers" and "Educational English Audio Data Recorded by L1 Korean Speakers," which include all read speech samples, are used. Table 1 summarizes the dataset statistics and surface-meaning divergence distribution. The Korean dataset contains 41,803 samples, while the English dataset contains 72,022 samples. Surface-meaning divergence is measured by the Levenshtein edit distance (ED) between surface and meaning transcriptions, with character-level syllables for Korean and word-level tokens for English. Both languages exhibit similar divergence distributions, with the majority of samples falling in the ED 0–3 range and only a small portion exceeding ED=10.

Table 1. Dataset statistics and Distribution of surface-meaning divergence using Levenshtein edit distance.

	Korean	**English**
Train	33,442	57,616
Validation	4,180	7,199
Test	4,181	7,207
Total	**41,803**	**72,022**
Surface–meaning edit distance		
ED = 0	16,203 (38.8%)	18,774 (26.1%)
ED = 1–3	14,790 (35.4%)	28,993 (40.3%)
ED = 4–10	8,617 (20.6%)	21,256 (29.5%)
ED ≥ 11	2,193 (5.2%)	2,999 (4.2%)

Baselines and Implementation Details. For SO, Whisper base and small (Radford et al., 2022) fine-tuned from pretrained weights and Conformer (Gulati et al., 2020) with a single transformer decoder are evaluated. For DO, a Conformer encoder with two separate transformer decoders is trained jointly on both objectives. The models are trained for 50 epochs using the AdamW optimizer. The weight decay is set to 0.01, and the learning rate is set to 10^{-4}. For fine-tuned models, the learning rate is reduced to 10^{-5}. The batch size is eight, and SpecAugment (Park et al., 2019) is used for data augmentation. Experiments are conducted on a single NVIDIA RTX 3090 GPU. The primary reported metric is the character error rate (CER), which is calculated using beam search decoding with a beam size of five.

[1]This research used datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. All data information can be accessed through AI-Hub (www.aihub.or.kr).

Table 2. Model performance reported in CER (%).

		Korean		English	
Model	**Params**	**Surface**	**Meaning**	**Surface**	**Meaning**
Single-output					
Conformer	32M	11.14	1.60	13.78	3.87
Whisper-base	72M	10.05	4.62	11.39	0.55
Whisper-small	244M	6.76	0.54	11.20	0.27
Dual-output					
Conformer	40M	11.34	0.77	15.08	3.19

3.2. Results

Table 2 shows the performance of Korean and English. For all models, surface transcription is consistently more difficult than meaning transcription. This pattern holds across both languages and all model scales. These results align with previous observations that verbatim pronunciation recovery is more challenging than retrieving the standard written form (Saraçlar & Khudanpur, 2004). Whisper scaling improves both transcription forms; however, the improvement is marginal for English, and the surface-meaning performance gap persists.

Joint MTL exhibits asymmetric effects across outputs. In both languages, MTL improves meaning transcription but degrades surface transcription. The degradation in surface transcription is substantially larger in English than in Korean, while improvements in meaning are comparable. This cross-lingual asymmetry under identical training conditions raises a central question addressed through stratified and representation-level analyses.

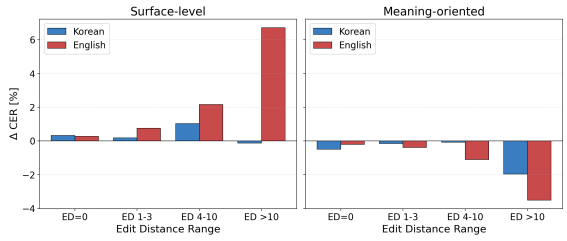


Figure 2. CER gap (Δ = DO − SO) stratified by ED across languages and outputs. Negative values indicate that DO improves over SO, positive values indicate degradation.

3.3. Analysis Beyond Distributional Differences

To examine whether cross-lingual asymmetry depends on surface-meaning divergence, the test sets are stratified by ED across two transcriptions. For each ED range, the CER gap is compared for surface and meaning outputs separately. Figure 2 shows the stratified results and reveals two patterns.

In Korean, MTL effects are minor and inconsistent with divergence. The surface gap ranges from +0.19 to +1.03 in the lower ranges and becomes slightly negative at ED>10. The meaning gap remains small, reaching -1.96 at ED>10. In English, the CER gap scales in opposite directions with divergence for the two outputs. The surface gap increases monotonically from +0.28 at ED=0 to +6.72 at ED>10. The meaning gap strengthens monotonically from -0.20 to -3.51, indicating a systematic surface-meaning trade-off that intensifies with divergence. Despite these starkly different MTL behaviors, the underlying ED distributions are similar across languages, as reported in Table 1. These findings suggest that the asymmetry cannot be attributed to distributional differences alone, thus motivating a representation-level analysis.

4. Mechanistic Analysis

To localize the source of the cross-lingual asymmetry in surface degradation, this section examines encoder representations and decoder representations with cross-output analysis. Centered Kernel Alignment (CKA) (Kornblith et al., 2019) is used to compare the representations of the SO and DO models across layers. This process reveals where the representations diverge or converge within the encoder and decoder.

Table 3. Layer-wise CKA between encoder representations.

	Korean			English		
Layer	S_{SO} ↕ M_{SO}	S_{SO} ↕ S_{DO}	M_{SO} ↕ M_{DO}	S_{SO} ↕ M_{SO}	S_{SO} ↕ S_{DO}	M_{SO} ↕ M_{DO}
0	0.95	0.98	0.96	0.91	0.70	0.88
3	0.43	0.72	0.51	0.89	0.88	0.88
6	0.56	0.74	0.60	0.75	0.72	0.84
9	0.45	0.78	0.48	0.66	0.67	0.70
11	0.56	0.81	0.62	0.40	0.46	0.53

4.1. Encoder Representation

Table 3 reports the CKA between the SO and DO conformer encoders at each layer. S and M refer to the surface and meaning outputs, respectively. S_{SO} refers to the surface SO encoder, M_{SO} refers to the meaning SO encoder, and S_{DO} and M_{DO} refers to the DO encoder viewed through the surface and meaning outputs, respectively. Notation X ↔ Y indicates the CKA between two representations X and Y.

In Korean, S_{SO}↔M_{SO} diverges sharply from Layer 3 onward, indicating that each encoder trains distinct representations. For the same output, S_{SO}↔S_{DO} remains consistently high in deeper layers, while M_{SO}↔M_{DO} shows moderate alignment. In English, however, the pattern differs. S_{SO}↔M_{SO} decreases gradually before dropping sharply at the final layer, suggesting that the two SO encoders learn similar representations. Only at the final layer does it drop

Table 4. Layer-wise CKA between decoder representations.

	Korean					English				
	Baseline	*Same-output*		*Cross-output*		*Baseline*	*Same-output*		*Cross-output*	
Layer	S_{SO} ↕ M_{SO}	S_{SO} ↕ S_{DO}	M_{SO} ↕ M_{DO}	S_{SO} ↕ M_{DO}	M_{SO} ↕ S_{DO}	S_{SO} ↕ M_{SO}	S_{SO} ↕ S_{DO}	M_{SO} ↕ M_{DO}	S_{SO} ↕ M_{DO}	M_{SO} ↕ S_{DO}
0	0.82	0.87	0.84	0.75	0.83	0.34	0.58	0.43	0.38	0.36
3	0.52	0.75	0.85	0.52	0.56	0.52	0.72	0.59	0.49	0.54
4	0.49	0.74	0.86	0.46	0.52	0.54	0.73	0.56	0.48	0.57
7	0.53	0.70	0.88	0.46	0.57	0.39	0.52	0.24	0.17	0.44

to 0.40, at which point the DO encoder also shows reduced similarity with the two SO encoders.

Since they produce different outputs, divergence of S_{SO}↔M_{SO} at the final layer was expected. However, pairs that target the same output also exhibit a similar drop in similarity, except for Korean S_{SO}↔S_{DO}. This suggests that, despite targeting the same output as the SO counterpart, the DO encoder fails to develop distinct representations. We refer to this phenomenon as "encoder-level entanglement." It is inconsistent with the meaning improvement observed in Table 2, which motivates analyzing how these representations are processed at the decoder level.

4.2. Decoder Representation

Table 4 shows the CKA between the SO and DO decoders, which are organized into three groups. The S_{SO}↔M_{SO} comparison illustrates how the two SO decoders compare. Same-output comparisons (S_{SO}↔S_{DO} and M_{SO}↔M_{DO}) compare each DO decoder with the SO decoder targeting the same output. The cross-output comparisons (S_{SO}↔M_{DO} and M_{SO}↔S_{DO}) instead pair each DO decoder with the opposing SO decoder.

In Korean, the baseline S_{SO}↔M_{SO} diverges from 0.82 in the input layer to 0.53 in the final layer. This shows that the two SO decoders develop distinct representations. Same-output S_{SO}↔S_{DO} decreases gradually with depth, while the M_{SO}↔M_{DO} approaches 0.88 at deeper layers. Same-output values consistently exceed cross-output values across all layers, indicating that Korean DO decoders align with their respective outputs.

However, the pattern is different in English. The baseline S_{SO}↔M_{SO} remains low from the beginning, peaking only in the middle layers before dropping in the final layer. The same-output S_{SO}↔S_{DO} increases in the middle layers, then falls to 0.52 at Layer 7. Meanwhile, M_{SO}↔M_{DO} drops to 0.24. Cross-output S_{SO}↔M_{DO} drops even further, reaching 0.17. These results confirm that M_{DO} constructs a representation distinct from all other decoders, effectively bypassing the entangled encoder. S_{DO} lacks this flexibility because it must remain connected to the encoder to produce frame-aligned transcription. This asymmetric flexibility explains why English exhibits significant surface degradation alongside moderate meaning improvement under MTL.

Overall, Korean decoders develop representations specific to the output, consistent with the encoder-level separation. The meaning decoder strengthens alignment beyond what the encoder achieves. However, the entangled English encoder provides no such foundation, limiting decoder-level adaptation. These findings suggest that encoder-level separation is a prerequisite for effective decoder-level specialization.

5. Conclusion

This paper investigated joint MTL for DO L2 ASR across Korean and English. MTL improves meaning but degrades surface transcription. The degree of degradation is substantially larger in English, as surface-meaning divergence increases. CKA analysis reveals that the Korean encoder learns disentangled representations, whereas the English encoder exhibits entanglement. At the decoder level, Korean decoders build on this separation, further strengthening output-specific alignment. However, in English, the meaning decoder compensates by constructing a distinct representation that bypasses the entangled encoder, while the surface decoder cannot. These findings suggest that encoder-level separation is a prerequisite for effective decoder-level specialization, motivating MTL approaches that mitigate encoder-level entanglement in DO L2 ASR. Promising directions include sparse decomposition, adversarial training, and gating mechanisms, along with validation through complementary similarity metrics beyond CKA.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25492127).

References

Ernestus, M. and Warner, N. An introduction to reduced pronunciation variants. *Journal of Phonetics*, 39(SI):253–

260, 2011.

Eskenazi, M. An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844, 2009. doi: 10.1016/j.specom.2009.04.005.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech*, pp. 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015. URL https://www.isca-archive.org/interspeech_2020/gulati20_interspeech.html.

Kim, S., Hori, T., and Watanabe, S. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4835–4839. IEEE, 2017.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Le, H., Pino, J., Wang, C., Gu, J., Schwab, D., and Besacier, L. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3520–3533, 2020.

Munro, M. J. On the difficulty of defining "difficult" in second-language vowel acquisition. *Frontiers in Communication*, 6:639398, 2021. doi: 10.3389/fcomm.2021.639398.

Nozaki, J. and Komatsu, T. Relaxing the conditional independence assumption of ctc-based asr by conditioning on intermediate predictions. *arXiv preprint arXiv:2104.02724*, 2021.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, pp. 2613–2617. ISCA, September 2019. doi: 10.21437/interspeech.2019-2680. URL http://dx.doi.org/10.21437/Interspeech.2019-2680.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Saraçlar, M. and Khudanpur, S. Pronunciation change in conversational speech and its implications for automatic speech recognition. *Computer Speech & Language*, 18(4):375–395, 2004. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2003.09.005. URL https://www.sciencedirect.com/science/article/pii/S0885230803000548.

Shakeel, M., Sudo, Y., Peng, Y., Lin, C.-J., and Watanabe, S. Unifying diarization, separation, and asr with multi-speaker encoder. *arXiv preprint arXiv:2508.20474*, 2025.

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., and Hayashi, T. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253, 2017.