ReShape Attention による音声と言語の基盤モデルの統合

叶高朋 [1]　小川厚徳 [1]　デルクロア・マーク [1]　チェン・ウィリアム [2]
福田りょう [1] 松浦孝平 [1] 芦原孝典 [1] 渡部晋治 [2]
[1] 日本電信電話株式会社　[2] カーネギーメロン大学
{takatomo.kanou, atsunori.ogawa, marc.delcroix, }@ntt. com
{ryo.fukuda, kohei.matsuura, takanori.ashihara}@ntt. com

概要

本論文では, 音声翻訳システムにおいて, 音声基盤モデルである Whisper の分散表現を言語基盤モデルである LLaMA2 の分散表現と結合する ReShape Attention (RSA) を提案する. RSA は, LLaMA2 の Transformer 層の内部に挿入され, 音声とテキストの分散表現を, 同じ特徴次元のサブベクトルに変形し, 2 つの分散表現間で Cross-Attention を実行する. RSA により, LLaMA2 と Whisper の勾配グラフは接続され, 音声翻訳システム全体を入力音声に対して最適化できる. RSA は, Whisper と LLaMA2 を用いた Cascade 音声翻訳システムと比較して, BLEU スコアを相対的に 8. 5% 向上させた. さらに, RSA は正解の書き起こしが得られる場合において Cascade システムよりも音声翻訳精度を向上させる可能性があることが示された.

1　はじめに

基盤モデルは, 大規模なデータで学習された巨大なモデルで, プロンプトに基づいて, さまざまな異なるタスクを高い精度で実行できる [1, 2, 3]. また, 音声翻訳のような専門的な下流タスクシステムを開発するための事前学習モデルとして利用することもできる. 音声基盤モデル Whisper と言語基盤モデル LLaMA2 を連結することで, 音声認識と機械翻訳の Cascade 型音声翻訳システムを構築できる. しかし, Cascade システムでは, Whisper による音声認識誤りの影響を受ける. 加えて, LLaMA2 は韻律や抑揚, 話者の性別など, 音声から得られる非言語情報を利用することができない. このような非言語情報は, 性差のある言語の翻訳やセグメントを学習するのに役立つことが知られている [4, 5]. 一方で, End-to-end (E2E) システムは入力音声を直接翻訳する. そのため, 音声認識誤りに頑健で, 非言語情報も学習することが可能である. しかし, E2E モデルは多くの原言語音声と目的言語テキストのペアデータを学習に必要とする [6, 7]. 第 3 の音声翻訳システムとして Hybrid システムがある. Hybrid システムは, Cascade と E2E の両方の長所を得ることを目的とし, 音声と書き起こしの両方を入力として翻訳を行う. 先行研究では, 音声基盤モデルと言語基盤モデルを結合する Cross-Attention 型の接続アダプターが提案された [8]. この接続アダプターは, 学習可能なパラメターを持ち, 基盤モデルの各層に挿入される. そのため, 基盤モデルが多くの層数を持つ場合, 接続アダプターのパラメータ数も膨大となる. そして, 多くのパラメータを基盤モデルに挿入することとなり, 破滅的忘却の発生や学習を難しくする可能性がある.

本研究では, Hybrid システムに焦点を当て, 先行研究の問題を解決するために ReShape Attention（RSA）を提案する. RSA は異なる次元数の分散表現を写像を行うことなく統合する. 初めに, RSA では音声基盤モデルと言語基盤モデルの分散表現を同じ次元数の Sub-vector へ分解する. そして, Sub-vector レベルで Cross-Attention を計算し, 2 つの分散表現を統合する. この時, RSA は写像を行わないため学習可能なパラメータを持たない. ゆえに, 基盤モデルの層数が増加しても, 学習可能なパラメータ量を増やすことなく基盤モデルを結合可能である.

本研究では, RSA の有効性を音声翻訳タスクで検証した. 音声基盤モデルには Distil-Whisper [9] を, 言語基盤モデルには LLaMA2-7B [10] を利用した. 音声翻訳のデータセットは MuST-C [11] を用いた. RSA は Cascade システムや, 先行研究の接続アダプターを持ちいた E2E, Hybrid システムよりも少ないパラメータにも拘わらず, BLEU スコアを相対的に 8. 5% 向上させた. さらに, Cascade システムと Hybrid システムにおいて仮に正しい音声書き起こしが得られた場合の性能を比較した. 分析の結果, 接続アダプター

が音声認識誤りの軽減のみならず, 非言語情報の学習にも有効であることを確認できた.

2 音声翻訳システム

本研究では, 後段の言語処理器への入力が異なる 3 つの音声翻訳システムを扱う. Cascade, E2E, Hybrid システムをそれぞれ図 1 に示す. Cascade システムは入力音声 X を変換し, 音声の分散表現 $\mathbf{E}^{\mathrm{x}} \in \mathbb{R}^{L\times D^{\mathrm{x}}}$ と, 書き起こしテキスト $Y \in \mathcal{V}^N$ を得る. ここで, L は分散表現の長さ, D^{x} は次元数, N はテキスト長, $\mathcal{V}$ は音声基盤モデルの辞書サイズを表す. そして, 言語基盤モデルは Y を入力として受け取り, 自身のトークナイザで再分割して翻訳を行う.

E2E システムでは, 音声の分散表現 $\mathbf{E}^{\mathrm{x}}$ を言語基盤モデルに入力し翻訳を行う. そのため, $\mathbf{E}^{\mathrm{x}}$ の次元数を言語基盤モデルの次元数 D^{y} を合わせる写像が必要である. 一般的には $D^{\mathrm{x}} \neq D^{\mathrm{y}}$ となる. この変換で得られる分散表現を接続分散表現 $\mathbf{E}^{\mathrm{o}} \in \mathbb{R}^{L\times D^{\mathrm{y}}}$ と呼ぶ. 言語基盤モデルは $\mathbf{E}^{\mathrm{o}}$ を入力として翻訳を行う. 本研究では, E2E システムの接続アダプターには Bottleneck 層 [12](図 2 (a)) , または, Transformer 層 [13] (SLM) [14] (図 2 (b)) を用いた. Bottleneck 層は ReLU 活性関数を持つ 2 層のニューラルネットワークである.

最後に Hybrid システム [8] では, 図 1 (c) に示すように, 音声の分散表現 $\mathbf{E}^{\mathrm{x}}$ とテキストの分散表現 $\mathbf{E}^{\mathrm{y}} \in \mathbb{R}^{N\times D^{\mathrm{y}}}$ を入力として受け取り内部で統合する. N はテキストの長さを表す. Radhakrishnan らは, Multi-head attention (MHA) を用いて次のように統合を行った [].

$$\mathbf{E}^{\mathrm{c}} = \mathbf{E}^{\mathrm{y}} + \underbrace{f(\mathbf{E}^{\mathrm{x}}, \mathbf{E}^{\mathrm{y}}; \theta)}_{\triangleq \mathbf{E}^{\mathrm{o}}} \in \mathbb{R}^{N\times D^{\mathrm{y}}}. \tag{1}$$

ここで, $\mathbf{E}^{\mathrm{c}}$ は, 統合された分散表現を表す. $f(\cdot)$ は, 統合機構を表しパラメータ θ に基づいて, 異なる 2 つの分散表現を写像し次元数をそろえ, Cross-Attention を行う. そして, $\mathbf{E}^{\mathrm{c}}$ を出力層（FFN）で変換し, 言語の分散表現に加算して接続分散表現 $\mathbf{E}^{\mathrm{o}}$ を生成する [8].

MHA [8] では 統合機構 $f(\cdot)$ は以下のようになる,

$$\mathbf{E}^{\mathrm{o}} = \mathsf{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\mathbf{S}_1, \ldots, \mathbf{S}_H]\mathbf{W}^{\mathrm{o}}, \tag{2}$$

$$\text{where } \mathbf{S}_h = \mathsf{Attention}(\mathbf{Q}\mathbf{W}_h^{\mathrm{Q}}, \mathbf{K}\mathbf{W}_h^{\mathrm{K}}, \mathbf{V}\mathbf{W}_h^{\mathrm{V}}), \tag{3}$$

ここで $\mathsf{Attention}(\cdot)$ は注意機構, $\mathsf{MultiHead}(\cdot)$ は MHA である. Query として ($\mathbf{Q} = \mathbf{E}^{\mathrm{y}}$) を, Key と Value として, それぞれ ($\mathbf{K} = \mathbf{E}^{\mathrm{x}}$, $\mathbf{V} = \mathbf{E}^{\mathrm{x}}$) を用いる. なお, $\mathbf{S}_h \in \mathbb{R}^{N\times D}$ は, MHA の第 h 番目の attention head, H は head 数を表し, $\mathbf{W}_h^{\mathrm{Q}} \in \mathbb{R}^{D^{\mathrm{y}}\times D}$, $\mathbf{W}_h^{\mathrm{K}}, \mathbf{W}_h^{\mathrm{V}} \in \mathbb{R}^{D^{\mathrm{x}}\times D}$, $\mathbf{W}^{\mathrm{o}} \in \mathbb{R}^{D^{\mathrm{y}}\times D^{\mathrm{y}}}$ は, それぞれ Query, Key, Value, 統合分散表現の写像重みを表す. これらの写像行列 $\mathbf{W}_h^{\mathrm{Q}}$, $\mathbf{W}_h^{\mathrm{K}}$, $\mathbf{W}_h^{\mathrm{V}}$ は, 音声と言語の分散表現の次元数を共通の次元数 $D = D^{\mathrm{y}}/H$ に揃える役割を担う.

MHA アダプターの写像行列 $\mathbf{W}_h^{\mathrm{Q}}$, $\mathbf{W}_h^{\mathrm{K}}$, $\mathbf{W}_h^{\mathrm{V}}$ は, ランダムに初期化され, 音声と言語の分散表現 ($\mathbf{E}^{\mathrm{x}}$ と $\mathbf{E}^{\mathrm{y}}$) は学習の初期段階においてランダムに写像されるため大幅な変更を受ける. このようなにランダムな写像によって変換された分散表現は, 学習を不安定にし, 学習初期の損失値を大きく増加させる. そのため不必要に基盤モデルのパラメターを更新し破滅的忘却を引き起こす可能性がある. 加えて, Cross-Attention 型の接続アダプターは基盤モデルの各層に挿入されるため, 基盤モデルの大きさに比例して追加される学習パラメータが多くなり, 学習が困難となる. これは, 年々巨大化する基盤モデルにおいて, 深刻化な問題となりうる [15].

3 ReShape attention による接続

本研究で提案する ReShape attention（RSA）は, 音声の分散表現の情報を写像などで破壊することなく, そのまま言語基盤モデルに受け渡すことを目標とする. それにより, 言語基盤モデルは, テキストの言語情報と音声の非言語情報の両方を考慮した音声翻訳を学習可能となる. そこで, 写像行列を持たない接続アダプターを提案する. 統合機構 $f(\cdot)$ (式 (1)) は, 音声の分散表現をテキストの分散表現と同じ長さ次元数に整形する. この時, 本研究では写像を行うことなく, 時間・次元の両方向のアライメントを同時に行う. この点は, 次元数の違いを写像行列 $\mathbf{W}_h^{\mathrm{Q}}$, $\mathbf{W}_h^{\mathrm{K}}$, $\mathbf{W}_h^{\mathrm{V}}$ で解決する先行研究と違う点である.

まず, 初めに 2 つの分散表現を $\mathbf{E}^{\mathrm{x}}$ と $\mathbf{E}^{\mathrm{y}}$ を共通の次元数 $\bar{D}$ に変形する. Reshape 機構 $\mathsf{RS}(\cdot)$ は以下のように動作する.

$$\mathsf{RS}(\mathbf{E}^{\mathrm{x}}; \bar{D}) \in \mathbb{R}^{LH^{\mathrm{x}}\times\bar{D}}, \quad \mathsf{RS}(\mathbf{E}^{\mathrm{y}}; \bar{D}) \in \mathbb{R}^{NH^{\mathrm{y}}\times\bar{D}}, \tag{4}$$

ここで, 共通の次元数 $\bar{D}$ は D^{x} と D^{y} の最大公約数であり,

$$H^{\mathrm{x}} = D^{\mathrm{x}}/\bar{D}, \quad H^{\mathrm{y}} = D^{\mathrm{y}}/\bar{D}. \tag{5}$$

となる. そして, Reshape された分散表現に対して Attention の計算を行う.

$$\bar{\mathbf{E}}^{\mathrm{o}} = \mathsf{Attention}(\mathsf{RS}(\mathbf{E}^{\mathrm{y}}; \bar{D}), \mathsf{RS}(\mathbf{E}^{\mathrm{x}}; \bar{D}), \mathsf{RS}(\mathbf{E}^{\mathrm{x}}; \bar{D})). \tag{6}$$

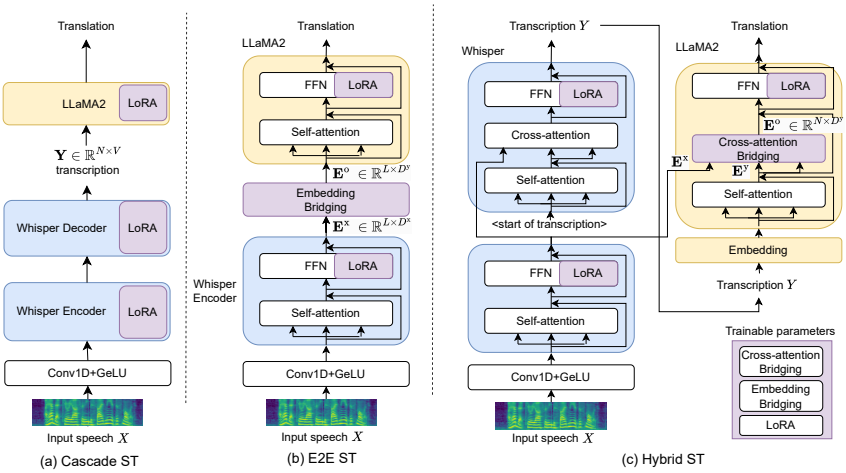

図 1　音声翻訳システムの概要図.

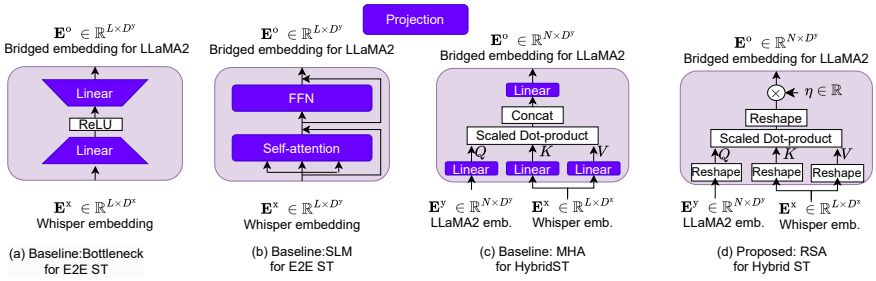

図 2　接続アダプターの概要図.

式 (6) の Attention(·) により, 接続分散表現 $\bar{\mathbf{E}}^{\mathrm{o}}$ を変換されていない元の $\mathbf{E}^{\mathrm{x}}$ と $\mathbf{E}^{\mathrm{y}}$ を用いて生成する. 最後に, $\bar{\mathbf{E}}^{\mathrm{o}}$ を $(N \times D^{\mathrm{y}})$ に変形し直して, 式 (1) を以下のように適用する.

$$\mathbf{E}^{\mathrm{o}} = \eta \mathrm{RS}^{-1}(\bar{\mathbf{E}}^{\mathrm{o}}; D^{\mathrm{y}}) \in \mathbb{R}^{N \times D^{\mathrm{y}}}. \tag{7}$$

ここで, $\mathrm{RS}^{-1}(\cdot; D^{\mathrm{y}})$ は Reshape 機構 $\mathrm{RS}(\cdot; \bar{D})$ を逆変換を表す. $\eta \in \mathbb{R}$ は, 学習可能なスカラーパラメーターで, 異なる分散表現のダイナミックレンジを調整する役割を担い, 初期値は 0 である. MHA と比べ, ReShape Attention (RSA) は写像行列を用いず次元数の違いを吸収する. そして, $\eta \in \mathbb{R}$ により, 段階的に音声の分散表現を加算していくため学習開始時点の損失値を Cascade システムと同様に抑え学習を安定させることができる.

4　実験

本研究では, MuST-C コーパスを用いて英独音声翻訳実験を行った [11]. MuST-C に含まれる既存のテストセットである “tst-HE” と “tst-COMMON” を結合して “Original” テストセットとした. また MuST-C よく知られた音声翻訳データであり, TED Talk から作成されているため, WEB から学習データを収集する基盤モデルにおいて, すでに学習データとして漏洩している可能性が高い. そこで, より厳密に評価を行うため, 基盤モデルリリース後に公開された TedTalk 1. 5 時間を用いて “New” テストセットを作成して評価に用いた. [1)]

4.1　モデル設計

モデル作成は ESPnet [16] を用いた. 音声基盤モデルとして Distill-Whisper を, 言語基盤モデルとして LLaMA2-7B を使用した. 音声基盤モデルと言語基盤モデルの分散表現の次元数はそれぞれ $D^{\mathrm{x}} = 1280$ と $D^{\mathrm{y}} = 4096$ である. 音声翻訳システム全体のパラメーターサイズは 7. 63B となる. 計算コストを減らすため 4bit の量子化を行い半精度計算を行った [17]. すべての音声翻訳システムは, FFN に LoRA アダプターを挿入して適応学習される [18]. LoRA アダプターの r と α はそれぞれ 64 と 128 に設定した.

初めに, 3 つの音声翻訳システム（図 1, Cascade, E2E, Hybrid. ）を構築した. Cascade システムは接続アダプターを持たず, LoRA アダプターのみで学習される. Whisper は音声を入力として, 書き起こしを正解として, LLaMA2 は正解書き起こしを入力と

1)　論文発表後 TalkList を公開予定

表 1　音声翻訳精度.

No.	System	Inputs for LLaMA2	Data	WER↓	BLEU↑	MTR↑	Trainable Params.↓
1	Cascade	Text	Original	7.7	25.4	36.1	119.3M
			New	11.6	23.1	35.3	
End-to-end Speech translation							
2	Bottleneck	Speech	Original	8.4	21.9	32.5	123.2M
			New	12.8	16.5	26.0	
3	SLM	Speech	Original	8.5	22.2	33.0	132.5M
			New	12.3	22.0	31.5	
Hybrid Speech translation							
4	MHA^{1st}	Speech & Text	Original	7.1	25.8	36.7	136.7M
			New	12.3	22.0	31.5	
5	MHA^{All}	Speech & Text	Original	7.5	25.5	36.0	564.9M
			New	12.0	23.0	34.8	
6	RSA^{All}	Speech & Text	Original	**6.8**	**26.6**	**37.1**	119.3M
			New	**9.7**	**24.3**	**35.8**	

して, 翻訳テキストを正解としてそれぞれ学習される. E2E システムは音声を入力として, 翻訳テキストを正解として学習される. Hybrid システムでは, 音声と正解書き起こしを入力とし, 書き起こしと翻訳テキストを正解とするマルチタスク学習が行われる. マルチタスク学習の損失は音声認識と翻訳でそれぞれ 0. 3 と 0. 7 で重みづけされる. Hybrid システムの接続アダプターでは, MHA と RSA の 2 種類のアダプターを用いた. まず, ベースラインとして MHA 接続アダプターを, LLaMA2 の Embed 層の後に 1 層のみ差し込んだ MHA^{1st} と, 32 層すべてに差し込んだ MHA^{All} を用意した. 接続アダプターは, Cross-Attention として動作し Whisper と LLaMA2 の分散表現を統合する. 提案手法である RSA ベースのシステムも同様に作成する. RSA^{All} は 32 層すべてに接続アダプターを挿入した (3 節, 図 2 (d)) . すべてのシステムは, NVIDIA A100 GPU で 12 時間適応学習され, 最も高い Validation スコアを達成したモデルで評価した.

4.2　音声翻訳結果

音声翻訳実験において, “Original” と “New” 評価セットの結果を表 1 に示す. 翻訳精度は SacreBLEU (BLEU) [19, 20], METEOR (MTR) [21], COMET [22] の 3 スコアで評価した. また, 各システムの音声認識率 (WER) も併せて報告する.

表 1 が示すように提案手法である RSA^{All} (system 6) が他のシステムより良い結果を示した. また, RSA^{All} は, 追加の学習パラメータがないにもかかわらず, 他のすべてのシステムに勝る結果を示した. これは, 多くの追加学習パラメータ（446MB）を持つ MHA^{All} と比べても, 性能的に優位であり, 基盤モデルに挿入された LoRA アダプターのパラメターのみでも十分な学習が行えることを示唆している. 加えて, RSA^{All} は, Cascade システム (system 1) にも勝っており, 精度と効率性の両方で他のシステムに優れることが分かった.

表 2　正解書き起こしを得られた場合の音声翻訳精度.

No.	System	Input	Data	BLUE↑	MTR↑	COMET↑
1	Cascade	Text	Original	26.7	37.2	81.5
			New	25.8	36.1	79.3
7	$\text{RSA}^{\textit{All}}$	Speech & Text	Original	**28.8**	**40.2**	**83.3**
			New	**28.1**	**40.0**	**83.0**

4.3　正解書き起こしによる翻訳結果

上記の実験での音声認識精度は, 90 %近い高精度を実現している. この実験では, 将来的に音声認識精度がさらに向上し, 仮に人と同じ認識精度を実現できたときに, 提案法が有効かどうか検証する. Cascade と Hybrid システムへの入力が音声と, その正解書き起こしだった場合に翻訳精度を比較した. 表 2 は, 正解書き起こしを得られた場合の翻訳精度を示す. 提案法の RSA^{All} は, 仮に正解書き起こしを得られた場合でも Cascade システムに勝る精度を示した. これは今後, 音声認識精度が向上したとしても Cascade システムに対して優位性を保てることを示している. また, 提案法が Cascade システムに勝る要因として, Hybrid システムは音声の分散表現を利用でき, 非言語情報を考慮した音声翻訳ができることが考えられる. これにより, Cascade システムよりも学習できる情報が増え, 翻訳精度の改善につながった.

4.4　おわりに

本研究では基盤モデル Whisper と LLaMA2 を統合した音声翻訳システムを構築した. 提案手法の RSA は追加パラメター無しで, 二つの基盤モデルを接合・全体最適化できる. 音声認識誤りに対する頑健性を向上させるのみならず, 非言語情報を音声翻訳に活用することで精度が向上する可能性を示した. 今後の展望として, より多くのドメインやタスクを通して提案法の有効性を確認する.

参考文献

[1] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[2] Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. Seamless: Multilingual expressive and streaming speech translation. **arxiv:2312.05187**, 2023.

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Miko l aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In **NeurIPS**, Vol. 35, pp. 23716–23736, 2022.

[4] Jie Jiang, Zeeshan Ahmed, Julie Carson-Berndsen, Peter Cahill, and Andy Way. Phonetic representation-based speech translation. In **MTSummit**, 2011.

[5] William Chen, Takatomo Kano, Atsunori Ogawa, Marc Delcroix, and Shinji Watanabe. Train long and test long:leveraging full document contexts in speech processing. In **ICASSP**, pp. 13066–13070, 2024.

[6] Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In **ICASSP**, pp. 7180–7184, 2019.

[7] Matthias Sperber and Matthias Paulik. Speech translation and the end-to-end promise: Taking stock of where we are. In **ACL**, pp. 7409–7421, 2020.

[8] Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A. Kiani, David Gomez-Cabrero, and Jesper Tegnér. Whispering LLaMA: A cross-modal generative error correction framework for speech recognition. In **EMNLP**, pp. 10007–10016. ACL, 2023.

[9] Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. **CoRR**, Vol. abs/2311.00430, , 2023.

[10] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. **CoRR**, Vol. abs/2307.09288, , 2023.

[11] Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a multilingual speech translation corpus. In **NAACL-HLT**, pp. 2012–2017. ACL, 2019.

[12] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. **Trans. Mach. Learn. Res.**, 2022.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Proc. NeurIPS**, Vol. 30, , 2017.

[14] Mingqiu Wang, Wei Han, Izhak Shafran, Zelin Wu, Chung-Cheng Chiu, Yuan Cao, Nanxin Chen, Yu Zhang, Hagen Soltau, Paul K. Rubenstein, Lukas Zilka, Dian Yu, Golan Pundak, Nikhil Siddhartha, Johan Schalkwyk, and Yonghui Wu. SLM: bridge the thin gap between speech and text foundation models. In **ASRU**, pp. 1–8, 2023.

[15] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. **CoRR**, Vol. abs/2403.14608, , 2024.

[16] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. ESPnet: End-to-end speech processing toolkit. In **Interspeech**, pp. 2207–2211. ISCA, 2018.

[17] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In **NeurIPS**, Vol. 36, pp. 10088–10115, 2023.

[18] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In **ICLR**. OpenReview.net, 2022.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In **ACL**, pp. 311–318. ACL, 2002.

[20] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation**, pp. 186–191, Belgium, Brussels, October 2018.

[21] Alon Lavie and Abhaya Agarwal. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In **WMT**, pp. 228–231. ACL, 2007.

[22] Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In **EMNLP**, pp. 2685–2702. ACL, 2020.