言語情報とパラ言語情報を考慮したニューラル音声翻訳

徳山 太顕[1]　Sakriani Sakti[1,2]　須藤 克仁[1,2]　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
{tokuyama.hirotaka.ti9, ssakti, sudoh, s-nakamura}@is.naist.jp

1 はじめに

異なる言語話者同士のコミュニケーションを可能にし，言語の垣根を越える音声から音声への翻訳 (Speech-to-Speech Translation; S2ST) が注目されている．従来，音声翻訳は自動音声認識 (Automatic Speech Recognition; ASR)，機械翻訳 (Machine Translation; MT)，テキスト音声合成 (Text-to-Speech; TTS) の 3 つのシステムで順々に開発されてきた．

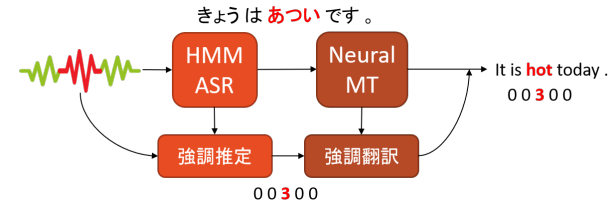
しかし，これらのシステムは言語情報の翻訳に焦点を当てており，感情や強調などの他の音声特徴を伝えるパラ言語情報を考慮していないものがほとんどである．最近の研究では，言語とパラ言語の認識と翻訳をそれぞれ別々にモデル化することでパラ言語の翻訳を提案している [1]．しかし，このシステムでは ASR が隠れマルコフモデル (HMM)，MT がニューラルネットワークと違うモデルをベースとしており，強調を推定，翻訳するシステムもそれぞれ ASR と MT の処理を待つ必要がある．したがって全体的に見れば複雑な構造をしており，翻訳システムとしても実装面においてもまだ最適ではない．

そこで本研究ではニューラルネットワークを用いることで言語とパラ言語の翻訳を統一的なモデルの中で考慮し，音声から強調を重視してパラ言語情報をテキストに反映できるかについて検討を行った．評価実験では自然音声データと音声合成データを用意し，それぞれパラ言語を適用できるようにデータを編集して学習させ，自然音声と音声合成でモデルの評価を行った．その結果，本研究で提案した手法は標準的な言語翻訳とほぼ同様の性能を維持したまま，言語情報とパラ言語情報の両方を翻訳できることが明らかになった．

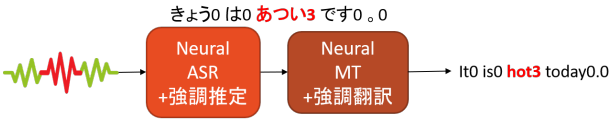
2 パラ言語情報の翻訳モデル

従来の翻訳システムにおいて，Do らはパラ言語情報を離散化して翻訳するシステム [1] を提案した．彼らのシステムは図 1(a) (従来の Cascade ASR-MT Model) のように，ASR 処理後に強調を推定，MT 処理後にその強調を適用している．例えば，“今日は暑いです．”という音声を，“暑い”を強調して入力して考えた場合，ASR では“きょう は あつい です”というテキストが出力され，その出力テキストを基に強調推定を行う．次にその“きょう は あつい です”というテキストを MT に入力することで“It is hot today .”と翻訳されるが，この時に日本語と英語の対応関係を表す Attention を利用して翻訳箇所と強調の関係を計算し，“hot”が強調されていることが分かるように出力される．しかし，これを一つの翻訳システムで考えた場合，様々なモデルを取り入れることによるシステムの大きさや複雑さ，さらに翻訳の遅延が発生するなど幾つかの制約がある．

そこで，今回の提案として図 1(b) (提案の Cascade



(a) 従来のパラ言語情報を利用した Cascade ASR-MT Model [1]



(b) 提案するパラ言語情報を利用した Cascade ASR-MT Model (例:Emphasis Tag)

(c) 提案するパラ言語情報を利用した Direct ST Multitask Model (例:Emphasis Tag)

図 1 パラ言語情報の翻訳モデル

ASR-MT Model) のように，ニューラルネットワークを利用して強調推定や強調翻訳を一つのシステムに統合することを考える．その場合，テキストデータと強調データを同時に翻訳させるため，テキストデータに強調データを付与させることを考える．

例えば，先ほどと同様の“今日は暑いです．”という音声を入力した場合，ASR では“きょう 0　は 0　あつい 3　です 0　。0”のように“あつい”に強調のデータを表すものを付与させる．次に，これを MT で翻訳した場合，“It0 is0 hot3 today0 .0” のように対応する単語に対して同様に強調のデータを出力できるように考える．強調を含むテキストデータについて，次の 3 種類のテキストデータを提案する．

Emphasis Tag:
きょう 0　は 0　あつい **3**　です 0　。0
It0 is0 hot**3** today0 .0

Emphasis 1-Token:
きょう　は　**<to3>**　あつい　です　。
It is **<to3>** hot today .

Emphasis All Token:
<to0> きょう <to0> は **<to3>** あつい <to0> です <to0> 。
<to0> it <to0> is **<to3>** hot <to0> today <to0> .

それぞれの ASR と MT を学習させた後，図 1(c) (提案の Direct ST Multitask Model) のように Ye らのマルチタスクモデル [2] をベースとした一つのモデルにし，音声からパラ言語情報を含む直接 Speech-to-Text (ST) 翻訳を行う．本来このマルチタスクモデルは ASR から，ASR，MT のデコーダを介さずに TTS のデコーダからスペクトログラムを出力するモデルであるが，今回は学習させた ASR のエンコーダと MT のデコーダを利用して学習を行わせる．つまり，同様に強調した音声を入力した際，“きょう 0　は 0　あつい 3　です 0　。0”を出力せず，そのまま MT のモデルを通して“It0 is0 hot3 today0 .0”と出力させることを考える．

また，言語によっては音声の強調を副詞等を利用して言語的に強調できる可能性がある [3]．したがって，もう一つの提案としてパラ言語情報を言語情報的に翻訳する実験も同時に行う．

3　データセット

3.1　使用するデータセットについて

本研究では日本語の母語話者による自然音声と，その書き起こしたテキストデータ，また，単語を挿入することでテキストの内容を強調したデータを利用する．これは 1029 文をそれぞれ 5 段階の強調

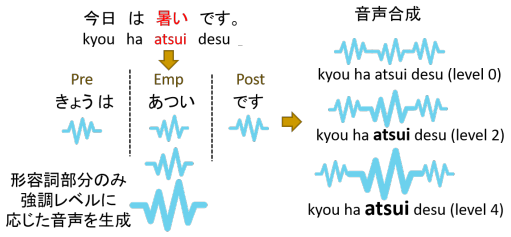


図 2　強調を含む音声合成の生成

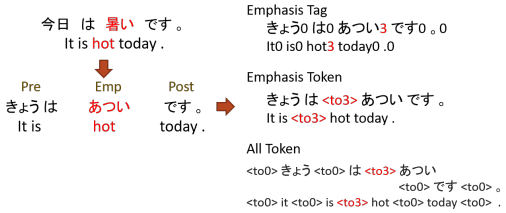


図 3　強調を含むテキストの作成 (強調レベル 3 の場合)

レベルで作成されたものであり，また，強調レベルに関しては 0 から 4 までのラベルが割り当てられ，その値が高いほど強調される単語が強く発音される．しかし，このデータは強調を含めて約 5000 文とデータセットとしては非常に少ない．そこで今回は BTEC (Basic Travel Expressions Corpus) コーパス [4] を利用する．ただし，BTEC には強調を含んだテキストデータや，音声データが含まれないため，これらは人工的に作成した．データ作成に関しては次の 3.2 節で述べる．

これらを利用して実験を行うが，前述のとおり自然音声のデータは BTEC と比べてはるかに少ないため，20 倍にオーバーサンプリングして学習データに追加して実験を行う．

3.2　強調を含む音声とテキストの生成

音声はまず，図 2 のようにテキストの一番初めに現れる形容詞の単語を中心に分離させる．次に各分離したテキストにおいて、Google Text-to-Speech[1] を利用して合成音声を生成させる．ここで音声の強調を反映させるように，形容詞部分の音声のみを音量を 5 段階に分けて変換させる．最後に各音声を連結させて一つの音声にすることで，強調を含む音声が生成できる．

また，図 3 のようにテキストも形容詞前後で分離させ，それぞれのシステムに合わせて情報を追加させた．

1)　https://pypi.org/project/gTTS/

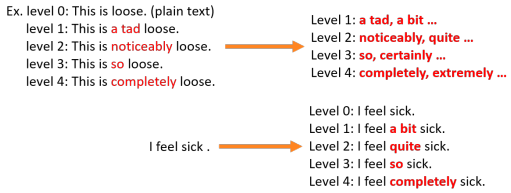


図 4　単語の挿入例

3.3　挿入を含むテキストの生成

まず自然音声のテキストデータのうち，単語が挿入されたテキストの中からどのような単語が挿入されているかを各強調レベルごとに抽出する．これらの抽出した単語のうち最もよく使われている単語 5 つをランダムに選び，図 4 のように形容詞の前に挿入してテキストデータを作成する．そこで MT を利用して，元のテキストデータから正しい位置に単語が挿入されるかを実験した．

4　評価実験

4.1　実験設定

実験はオープンソースの OpenNMT-py [5] を利用した．エンコーダとデコーダは Transformer [6] で構成し，最適化には Adam を使用した．このモデルで ASR と MT を別々に学習させてテストを行い，それぞれの結果から評価を行った．続いて ST 翻訳では ASR と MT のそれぞれの学習済みのモデルを利用して再度学習させ，MT と同様に評価を行った．また，単語を挿入する実験も同様に MT を利用した．

4.2　評価方法

評価方法として，まず各提案したデータセットが，ベースラインである一般的な ASR や MT (No Emphasis) と，2 節でも述べた従来法 (Separated) と比較しながら評価を行う．ASR では単語誤り率 (Word Error Rate : WER) を評価し，MT と ST 翻訳では BTEC と自然音声のテキストデータで，Multi-BLEU, SacreBLEU[2] を用いて評価する．

また，強調を含む翻訳システムにおいては，出力テキストを評価する言語評価と，強調部分が正しいかを評価する強調評価に分けて行う．それぞれの評価の際，強調を表すタグとトークンを分離させ，強調評価の場合は，全単語での強調レベルを正解データとの F1-score を計算する．また，挿入の実験では，出力が元のテキストを変化させず，各強調レベルに合った単語を挿入できている割合で評価した．

2)　https://github.com/mjpost/sacrebleu

4.3　実験結果

4.3.1　言語評価

言語評価は表 1 の通りである．Synthetic においてはトークンを利用したものが ST 翻訳において 2〜5 ポイントほど高いことが分かる．特に，All Token は，No Emphasis と同程度，もしくは 2 ポイント前後向上した．これはトークンを利用することで語彙数の変化が小さい事や，トークンを利用して強調レベルを学習できることが考えられる．

逆にタグを付けたものでは ASR の精度が少し向上する一方で，ST 翻訳のスコアは低下した．これは各単語にタグを付けることによって語彙数が増えることから，ASR においては強調が違えば別の単語であると認識できる一方で，MT や ST 翻訳においては別の単語であることから翻訳モデルを複雑にしてしまうことが原因であると考えられる．

一方，Natural での ST 翻訳の評価はかなり悪くなってしまった．これは ASR での影響が大きいため，自然音声の学習データを増やすなど ASR のスコアを改善する必要がある．

4.3.2　強調評価

強調評価は表 2 の通りである．ASR においては従来法よりも同等もしくは高いスコアを出しており，提案法が比較的優れている．特に提案法では，言語評価を含めて音声と編集したテキストデータで学習できていることが分かる．

また，MT や ST 翻訳においては，強調の含まれる単語の位置が翻訳結果に依存するため，文レベルにおいての強調は正しいとしても，単語の位置による強調が違うことが多かったため，ASR と比べて値は低い．しかし，提案法においては従来法よりも値は大きく，強調を含めた翻訳においても提案法が優れていると言えるであろう．また，従来法での値が非常に低い理由として，強調が Attention の関係性に依存する一方で，該当する単語に強調が当たってない結果が多かったことが挙げられる．また，翻訳結果によって表現が変わることも考えた場合，別の評価方法も考慮する必要がある．

4.3.3　挿入評価

挿入評価は表 3 の通りである．1-Token の方がより高い精度で挿入を行えていることが分かるが，これは各トークンと挿入する単語が対応しているこ

表 1　言語評価

Test Data	System		Cascade ASR-MT Model			Direct ST Multitask Model	
			ASR	MT			
			WER↓	Multi-BLEU↑	SacreBLEU↑	Multi-BLEU↑	SacreBLEU↑
Synthetic	従来法	No Emphasis	0.72	44.84	47.80	33.88	36.88
		Separated	0.38	43.99	46.71	—	—
	提案法	Emphasis Tags	0.64	42.64	45.57	29.24	32.57
		Emphasis 1-Token	0.80	42.31	45.64	37.92	41.29
		Emphasis All Token	0.80	44.84	47.71	35.94	39.37
Natural	従来法	No Emphasis	24.72	31.81	34.68	2.72	4.10
		Separated	19.50	28.74	31.76	—	—
	提案法	Emphasis Tags	24.37	34.12	36.54	2.08	2.34
		Emphasis 1-Token	25.45	36.05	38.71	5.18	6.17
		Emphasis All Token	22.40	34.35	36.91	5.29	6.75

表 2　強調評価 (F1-score)

Test Data	System		Cascade ASR-MT Model		Direct ST Multitask Model
			ASR	MT	
Synthetic	従来法	Separated	95.52	5.41	—
	提案法	Emphasis Tags	95.38	64.52	55.74
		Emphasis 1-Token	98.51	70.77	66.67
		Emphasis All Token	96.97	74.63	55.88
Natural	従来法	Separated	59.66	34.31	—
	提案法	Emphasis Tags	69.23	36.76	32.50
		Emphasis 1-Token	63.59	49.48	43.14
		Emphasis All Token	61.60	44.22	42.71

表 3　挿入評価

Test Data	System		Emphasis Level					Total
			0	1	2	3	4	
Synthetic	提案法	Emphasis Tags	98.18	87.88	93.94	94.94	90.91	96.81
		Emphasis 1-Token	99.20	90.91	100.00	100.00	100.00	98.90
		Emphasis All Token	94.33	93.94	96.97	90.90	93.94	94.25
Natural	提案法	Emphasis Tags	84.53	67.03	81.11	80.90	75.28	77.85
		Emphasis 1-Token	93.62	82.98	92.55	91.30	91.30	90.34
		Emphasis All Token	88.04	73.91	79.35	81.11	82.22	80.92

とが考えられる．一方で Tags や All Token の場合では精度が比較的低い．これらは全ての単語にタグやトークンが付いていることから，学習時にそれぞれの挿入される単語との関係が取れていないことや，それに関連して出現回数が少ない単語に対して対応できていないことが考えられる．

5　おわりに

本研究ではパラ言語情報をニューラルネットワークを用いて翻訳することを目的として，2 節のようなモデルやパラ言語情報を含めたテキストを提案した．続いて，3 節のようにパラ言語情報を含むデータセットを準備して実験と評価を行った．

実験結果から，従来法と比べると言語評価，強調評価の両方の観点から提案法が高いスコアを出すことができた．提案法においてはシステムによって得手不得手があるが，ST 翻訳においてはトークンを利用したものが優れた結果を出しており，実用化を考えた場合は Natural でも従来法より高いスコアを出している All Token が最も適していると考える．また，今回は音声データも含め，学習データの大半が BTEC のものであるため Natural における結果は優れているとは言えないが，Synthetic における評価から学習データによっては現在よりも精度の高いモデルに学習できる可能性がある．

今後の展望として，音声の強調レベルや挿入データに関して人間による主観評価を行うことや，言語強調についても直接翻訳が行えるようにモデル考えること必要がある．また，自然音声のデータを確保して Natural の精度を向上させることを目指す．

謝辞

本研究の一部は JSPS 科研費 JP17H06101 の助成を受けたものである．

参考文献

[1] Quoc Truong Do, Sakriani Sakti, and Satoshi Nakamura. Sequence-to-sequence models for emphasis speech translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 26, No. 10, pp. 1873–1883, 2018.

[2] Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model, 2019.

[3] Quoc Truong Do, Sakriani Sakti, and Satoshi Nakamura. Toward multi-features emphasis speech translation: Assessment of human emphasis production and perception with speech and text clues. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 700–706, 2018.

[4] Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. Creating corpora for speech-to-speech translation. In *Eighth European Conference on Speech Communication and Technology*, 2003.

[5] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. Opennmt: Open-source toolkit for neural machine translation, 2017.

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