入力音声の継続長を考慮した翻訳システム*

☆叶 高朋，Sakti Sakriani，Graham Neubig，戸田 智基，中村 哲 (奈良先端大)

1 はじめに

人の会話では発話内容だけでなく，表情，話し方，間の取り方，声の抑揚などの非言語情報も内容を理解するのに非常に重要である．このため，人手による音声翻訳・通訳において翻訳者は，このような視覚的，音声的な特徴から得られる非言語情報を加味して翻訳している [1]．一方，現在の音声翻訳は音声認識・機械翻訳・音声合成の 3 つのモジュールで構成され，各モジュールでは言語情報のみをやりとりしているため，話し手の表情や音声のもつ非言語情報を翻訳へ反映できない [2]．

そこで，本論文では音声に着目し，言語情報だけでなく音声から得られる非言語情報も同時に翻訳する音声翻訳を提案する．入力音声の音声特徴量 (F0, 継続長，スペクトル等) を翻訳音声上に再現し，自分が母国語で話しているかのような音声翻訳の実現が本研究の最終目標である．そのため，音声の違いを連続的に変換可能な音声特徴量変換モデルを設計した．本論文では様々な音声特徴量を扱うという目標の第一歩として継続長に着目した．本研究では，入力音声より HMM(隠れマルコフモデル) の各状態ごとに継続長を抽出し，目標音声の継続長との関係を表す変換行列を学習し翻訳に用いた．その結果，入力音声の継続長情報が翻訳音声に反映され，翻訳音声上の強調位置を予測する主観評価にて効果が確認された．

本論文の構成は，2 章で，先行研究を紹介し 3 章で，提案モデルの処理について述べ，4 章で比較実験を通して考察し，5 章で今後の課題と研究の方向性について述べる．

2 先行研究

機械翻訳に非言語情報を利用した研究例として，音声の抑揚の違いにより翻訳文の曖昧性を解消する研究 [3][4]，入力音声と正解音声の音響的類似性を利用して音声認識誤りに頑健な翻訳システムを構築する研究 [2] などがある．一方，提案手法では言語情報を翻訳するモデルと非言語情報を翻訳するモデルの二つがあり，音声の特徴を翻訳するモデルにより，入力音声の特徴を翻訳音声上に再現できることが先行研究とは異なる．

3 継続長を考慮した音声翻訳の構成

提案手法の構成は，音声認識による特徴抽出部，言語，非言語翻訳部・音声合成部によって構成される．既存の音声翻訳システムと同じ構成であるが，本研究では各モジュールで音声特徴量も扱う．本論文は，多数ある音声特徴量 (F0，継続長，スペクトル等) の翻訳の中でまず最初のフェーズとして継続長に着目し，継続長の翻訳モデルを設計した．本研究で音声翻訳に利用する情報は，入力音声の言語情報と継続長，目標音声の言語情報と継続長である．本研究では非言語情報の翻訳モデルが翻訳音声に与える影響に焦点を当てるため，テキスト翻訳が容易な英語数字と日本語数字の小規模語彙の翻訳タスクを扱った．

非言語情報の翻訳は次のように設計した．まず，音声からの特徴抽出のために HTK(Hidden Markov Model Tool Kit) を用いて単語ごとに各単語 32 状態の HMM 音響モデルを構築し，HMM 状態系列の継続長を特徴量として抽出した[1]．継続長は単語ごとの単語継続長ベクトルとして扱い，原単語の継続長ベクトル $\mathbf{x}$ から目標単語の継続長ベクトル $\mathbf{y}$ へと翻訳した．翻訳には下記の線形重回帰モデルを利用した．$\mathbf{W}$ はバイアス項を含む回帰行列である．

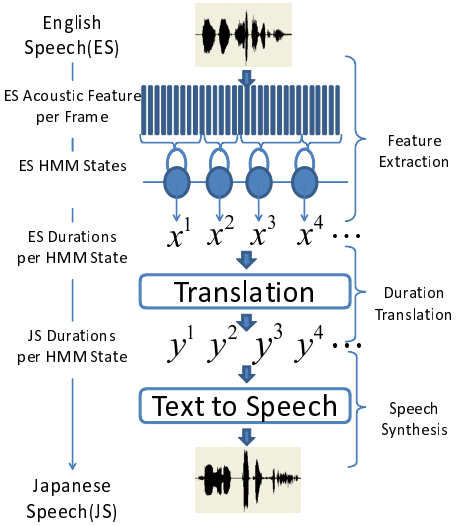

Fig. 1 提案モデルの概要図

$$\mathbf{y} = \mathbf{W}\mathbf{x}^T \tag{1}$$

モデルの回帰行列を学習するために，下記の評価尺度に基づき二乗平均平方根誤差を最適化した．

$$\arg\min_{\mathbf{W}} \sum_{n=1}^{N} ||\mathbf{t}_n - \mathbf{y}_n||^2 + \alpha\ ||\mathbf{W}||^2 \tag{2}$$

式 (2) において N はサンプルの総数，n はサンプル番号を表している．$\mathbf{t}$ は実際の日本語継続長のベクトルである．α は正則化項にかかる超パラメータである．

最後に音声合成は，コンテキスト情報を利用しない単語ベース音声合成を設計した．特徴抽出・翻訳と整合性を考慮し，音素ではなく単語単位の音声合成モデルを作成した．また，本研究では合成音声の継続長情報は翻訳部より与えられる．

4 評価・考察

評価実験では，海外との電話でチケット等の予約の確認時に聞き手がチケット番号を聞き間違え，話し手は相手が間違えた箇所を強調して言い直すシーンを想定した．この場合，聞き手は話し手の強調により間違い箇所を特定し訂正することが可能である．言い直しによる強調箇所の特定は言語情報だけでは困難である．本研究は上記の想定で日英二ヶ国語話者に数

[1]単語ごとに HMM を学習した理由は，継続長を翻訳は単語ごとに継続長翻訳を行う方が音素ごとにアライメントを考えモデル化するより容易であり，継続長の予測が容易になると考えたためである．

*A duration-sensitive speech translation system. by Takatomo KANO, Sakriani SAKTI, Graham NEUBIG, Tomoki TODA, Satoshi NAKAMURA

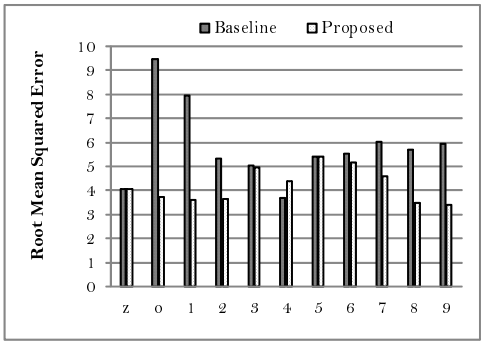

Fig. 2　横軸の各数字の翻訳モデルに対する客観評価

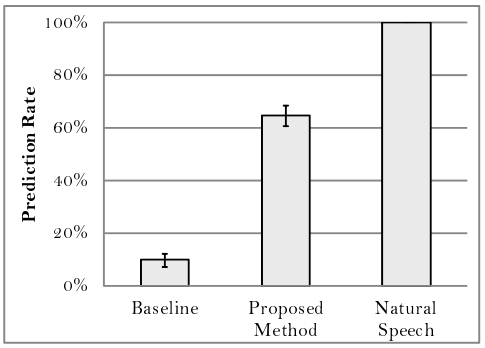

Fig. 3　強調単語認識率の評価

字列を読み上げ 1ヶ所強調して発話したコーパスを収録しモデル構築に利用した．音声認識における各種設定は，AURORA2[5] の論文を参考にした．発話内容は AURORA2 コーパスより GreedySearch[6] により獲得した単語バランス文 500 文である．収録したデータの分析により，強調音声は音声が長くなる傾向や強調の直前に長い無音区間ができるなど，音声の長さに関する変化が確認できた．このコーパスを用い，提案モデルを Baseline モデルと比較評価した．

提案モデル:発話された英語音声から得た継続長情報を翻訳し音声合成に用いるモデル

Baseline モデル:発話された日本語音声の平均的な継続長情報を音声合成に用いるモデル

また，これらの翻訳結果は言語情報が同等であり，非言語情報，継続長のみが異なっている．これらを，発話内の強調を認知できるかどうかについて比較した．提案モデル，Baseline モデルとも 1 話者 500 対の対訳音声コーパスのうち 445 文を学習データに用い，残り 55 文のうち 1 単語のみで構成される文を省き 53 文で評価した．

まず，客観評価として重線形回帰モデルにより英語継続長を翻訳した継続長が日本語の継続長との二乗平均平方根誤差が小さくなることを Fig.2 に示す．これにより，本提案モデルは平均的な日本語よりも目標音声の継続長の特徴を表現できている考察できる．

次に，翻訳音声に対する効果を 3 人の被験者に対し各手法 50 文づつ主観評価にて評価してもらい，翻訳音声上に強調情報が現れているか検証した．主観評価より，強調位置の予測 (Fig3) と強調の度合い (Fig4) を，

1:強調だと分からない

2:強調だと言われれば分かる

3:よく強調されている

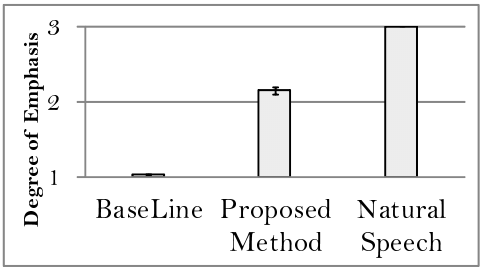

Fig. 4　強調の強度評価

の 3 段階評価した．Fig3 より，提案手法は Baseline を上回る 65%の高い認識率を示した．また，強調の強度も Baseline と比べ提案手法に優位な差があった．これらより手案手法が継続長翻訳により入力音声上の強調という非言語情報の翻訳に成功したと考察できる．また，詳しく分析した結果，提案手法では強調だと誤認識した単語の位置が正解の前後に偏る傾向が見られた．これは，提案モデルでは入力音声の継続長情報しか翻訳音声できず，入力音声のパワーなど特徴量が反映できていない．よって，被験者は音声の継続長のみで強調を判断せねばならず，発話において長い空白区間の前の単語を強調と感じるか後の単語を強調と感じるかの違いが現れたものだと考えられる．

5　まとめ

本研究は入力音声の継続長情報抽出し回帰モデルによって翻訳することで翻訳音声上に入力音声の音声的特徴を再現した．今後の研究として，同じ枠組みで話速を翻訳する研究や，他の音声特徴量に着目した研究，一般的な音声翻訳問題への拡張，コンテキスト情報を用いた合成音声の質向上などに取り組む．

6　謝辞

本研究は，(独) 情報通信研究機構の委託研究「知識言語グリッドに基づくアジア医療交流支援システムの研究開発」の一環として実施した．

参考文献

[1] S.Ogata *et al.*, Muti-Modal Translation System by Using Automatic Facial Image Tracking and Model-Based Lip Synchronization, Siggraph, p231, 2001.

[2] J.Jiang *et al.*, Phonetic Representation-Based Speech Translation, Proceedings of Machine Translation Sumit 13, page81-88, 2011.

[3] T.Takezawa *et al.*, A Japanese-to-English speech translation system:ATR-MATRIX, Spoken Language Processing, page957-960, 1998.

[4] W.Wahlster, Robust translation of sopntaneous speech:A multi-engine approach, Artificial Intelligence, page1484-1493, 2001.

[5] AURORA-2J, http://www.slp.cs.tut.ac.jp/CENSREC/data/AURORA-2J-data.pdf.

[6] J.S Zhang *et al.*, An Efficient Algorithm to Search For A Minimum Sentence Set For Collecting Speech Database, International Congress of Phonetic Sciences, page3145-3148, 2003.