一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2016-64（2016-12）

［招待講演］音声研究と自然言語研究の融合に向けて
～ 音声翻訳研究の過去と未来 ～

中村　哲†

†奈良先端科学技術大学院大学　情報科学研究科

〒630-0192 奈良県生駒市高山町 8916-5

E-mail:　†s-nakamura@is.naist.jp

あらまし　本発表では，音声翻訳を音声処理と自然言語処理の融合と捉え，これまでの研究，そして，今後の研究の方向性について考察する．

キーワード　音声翻訳，音声認識，機械翻訳，統計モデル

[Invited Talk] Towards Fusion of Speech and Natural Language Processing Research
－ Past and Future of Speech Translation Research －

Satoshi Nakamura†

† Graduate School of Information Science, Nara Institute of Science and Technology

8916-5 Takayama Ikoma, Nara, Japan,

E-mail:　†s-nakamura@is.naist.jp

Keywords Speech Translation, Speech Recognition, Machine Translation, Statistical Modelling

1. 音声翻訳研究のこれまで

1.1. 音声翻訳研究の創生期

自分の話した音声をその場で認識，翻訳，音声合成して相手言語で出力し，異なる言語を話す人々との自由なコミュニケーションを実現することは，人類の長年の夢である．我が国は世界に先駆けて 1986 年に国際電気通信基礎技術研究所（ATR）内に（株）ATR 自動翻訳電話研究所を発足させ，いち早くこの研究課題に着手した．1980 年以前の音声認識は信号処理，パターン認識，動的計画法などを基礎とし，音声合成は声道，音源モデルに基づく共振，反共振による音声合成フィルタ等を基礎としていた．一方で，自然言語処理は，文法理論，ロジックをベースとしており，異なる研究分野であった．音声翻訳はこれらのモジュールの統合技術であり，そういう意味で，ATR の音声翻訳プロジェクトは最初の音声研究者と自然言語処理研究者の共同作業の場であったと言える．筆者はこのプロジェクトにその初期から色々な形で関わってきた．本稿では，音声翻訳のこれまでと今後について，音声研究と言語研究の融合という視点で述べる．音声翻訳に関するより詳細な解説は［1,2］を参照されたい．

1.2. 音声認識における音響モデルと言語モデル

1980 年代に音声認識は情報理論をベースにした統計モデルの時代を迎える．隠れマルコフモデル（HMM）により，音声の発話毎の変動を正規分布の集合からの生起確率で表現でき，時間変動を状態遷移確率により表現できるようになった．さらに，単語列の確率を表現する N-gram 言語モデルを用いることで，連続発話音声の最尤復号という形で任意の単語列を認識できるようになった．それまで，オートマトンによる決定的な文法に基づく動的計画法が用いられていたが，音声認識全体が最尤復号の形で再定義され，確率を用いて音響モデルと言語モデルが統合されることになった．この時点が，音声処理と言語処理の最初の本格的な出会いであったと思われる．1990 年代に入り，ATR では N-gram からより広い文法のクラスを扱うことのできるシフトリデュースパーザを HMM に組み合わせた HMM-LR 音声認識が研究された[3]．さらに，1990 年代初期には第 2 期のニューラルネット（NN）研究が始まり，NN による言語モデルの最初の試みも行われた[4]．しかし，この時代は，音声研究者が N-gram, NN 等による統計的モデルにより言語情報を活用し始めた段

階であった.

1.3. 統計的機械翻訳による音声, 言語モデル

1990 年に IBM から統計的機械翻訳の論文が発表され，2002 年に多数の素性を用いるための Loglinear 法が，2003 年にフレーズベース統計翻訳が提案され，統計的機械翻訳のコアになる理論が確立された．この統計翻訳は，雑音のある通信路モデルに基づき，原言語の文を翻訳モデル，語順入替えモデル，対象言語の言語モデルにより同時に復号する．背景になる通信路モデルが同一であるため，音声認識と機械翻訳はほぼ同一の枠組みで行える．ATR は 2000 年からこの枠組への方針変更に取り組んだ．これにより，プロジェクト開始から 14 年を経過してやっと音声研究者と言語処理研究者が同じ言葉で議論が行える時代が到来した．尤度最大化，誤り最小化によるモデル学習等，同様の方法が適用できるようになった.実応用を目指し,ATR では対象分野を旅行会話と設定し必要な対訳コーパスの収集を行った．多様な話し言葉への対応，対訳文の多言語翻訳による言語対の拡張，実応用データの収集と改善のエコループの形成が可能となり，現在の音声翻訳の実用展開に繋がっている.

2. 今後の展開

2.1. 音声同時通訳

音声翻訳の究極の目標は，人間の同時通訳者に匹敵する同時通訳の達成である．現在の音声翻訳は，音声認識は入力音声に若干遅延して処理が可能ではあるが,候補の確定は発話区切りを確定してからでないとできず，また，機械翻訳は発話文が確定してからしか行えない．さらに，音声合成は機械翻訳出力が確定してからしか行えない．音声同時通訳のためには，音声のポーズ，基本周波数などの韻律などを用いた発話区切りの検出，構文構造の推定，逐次音声認識結果の出力が必要であり，さらには，言語間の文構造の違いを考慮した翻訳制御メカニズムが必要である［5]．これからの音声同時通訳には，パラ言語，言語情報，統語情報を融合，統合し，活用する必要がある.

2.2. パラ言語情報の必要性

人間が本来コミュニケーションに用いている情報は，言語情報（通常のテキスト情報）に加えて，パラ言語情報（韻律, 強調, 感情), 非言語情報（個人性等）がある．これまで，音声認識では如何に正確に音声を文字化するかが研究対象で，機械翻訳では汎用のテキスト翻訳の性能を改善し音声翻訳に適用するかが主眼であったため，パラ言語情報に注目することがなかった.しかし,コミュニケーションにおける情報伝達度，意図の伝達度を考慮すると，パラ言語情報が非常に重要な役割を担っている．今後はリアルタイムの音声コミュニケーションにおいて重要なファクターとしてパラ言語情報の研究を行う必要がある[6].

2.3. 音声翻訳と対話制御

これまで，音声翻訳はテキスト翻訳に音声入出力を付与したものと捉えられてきたが，旅行会話では入出力が異なる言語で行われる目的指向の対話と捉えることもできる．音声翻訳でもアプリケーションとしては目的が達成できたかどうかが最終の評価基準になるからである．特に，日本語は文脈依存言語であり，話し言葉ではよりその性質が強いため，対話中に主語省略や不完全文が多発する.

2.4. ニューラルネットと機械翻訳

Mikolov らにより分散表現が提案され，自然言語における単語表現が連続空間のベクトルとして取り扱えるようになった．これにより音声処理と全く同様なモデリングが利用できる．また，2014 年に Sutskever らによりニューラルネットに基づく機械翻訳（NMT）が提案された．畳み込みＮＮ（CNN)，時間遅れ NN（TDNN)，再帰型 NN, Long-Short Term Memory(LSTM) 等は既に知られていたが，機械翻訳に導入され大きな改善をもたらした．2015 年には Bahdanau らにより注意機構付き LSTM が提案され，現時点で最も高い性能を実現している．単語や文を連続値表現する方法の登場により，音声も自然言語もすべて同様のモデリングを導入できるようになりつつある.

3. まとめ

言語情報に加えてパラ言語情報を音声から抽出し，解析，変換することは非常に興味深い課題である．また，音声翻訳全体を一つのシステムとして扱えることができれば，翻訳精度に基づくモジュール最適化［7］や，End-to-end の強化学習なども可能になる．今後，音声処理と自然言語処理の研究全体を俯瞰した新しい挑戦的な研究が生まれることを期待する．多くの共同研究者，学生諸君にこの場を借りて深謝する.

文　献

[1] 中村　哲, “話し言葉の音声翻訳技術,” 電子情報通信学会誌, vol.96(11), pp865-873, 2013.11

[2] 中村　哲, ” 音声翻訳技術概観 “, 電子情報通信学会誌, vol 98(8), pp.702-709, 2015.08

[3] T.Hanazawa,et al., “ATR HMM-LR continuous speech recognition system”, Proc. ICASSP1990

[4] M.Nakamura, et.al., “A Study of English word category prediction based on neural networks”, Proc. ICASSP 1990

[5] T.Fujita, et al., “Simple, Lexicalized Choice of Translation Timing for Simultaneous Speech Translation,” Proc. of INTERSPEECH 2013

[6] Q.T.Do,et al., “Transferring Emphasis in Speech Translation Using Hard-Attentional Neural Network Models”,Proc. INTERSPEECH 2016

[7] M.Ohgushi,et al., “An Empirical Comparison of Joint Optimization Techniques for Speech Translation,” Proc. , INTERSPEECH 2013