2008－SLP－72 (15)
2008／7／19

同時通訳者の知識と韻律情報を用いた講演文章のチャンキング

清水 徹[†,††,‡] 中村 哲[†,††] 河原 達也[†,‡]

†独立行政法人 情報通信研究機構 知識創成コミュニケーション研究センター
††ATR 音声言語コミュニケーション研究所
‡京都大学情報学研究科

†,††〒619-0288 京都府「けいはんな学研都市」光台 2-2-2
‡〒606-8501 京都府京都市左京区吉田本町

E-mail: {tohru.shimizu, satoshi.nakamura}@nict.go.jp, kawahara@i.kyoto-u.ac.jp

あらまし 文の区切りが明確でない、一文が長くなる、文の途中に間(ポーズ)が空くなどの現象が見られる自然な話し言葉を、適切な単位に区分化する処理が求められている。本稿では、区分化の単位として従来用いられている文や節に代わる、プロの同時通訳者が原言語からターゲット言語に変換する自然なタイミングである音声翻訳単位を提案し、同単位の特徴と、言語情報ならびに韻律情報を用いた音声翻訳単位境界の推定手法について述べる。ポーズを伴う音声翻訳単位境界では人間の境界付与精度に近い推定精度が得られ、ポーズを伴わない音声翻訳単位境界の推定精度は低いものの、韻律情報に基づく素性の導入により、推定精度の向上効果が確認された。

Segmentation of spoken monologue using human interpreter's knowledge and prosodic features

Tohru SHIMIZU[†,††,‡] Satoshi NAKAMURA[†,††] and Tatsuya KAWAHARA[†,‡]

† Knowledge Creating Communication Research Center, National Institute of Information and Communication Technology
††ATR Spoken Language Communication Research Labs.
‡ School of Informatics, Kyoto University

† ‡ 2-2-2 Hikaridai, Keihannna Science City, 619-0288, Japan
‡〒606-8501 Sakyo-ku, Kyoto, 606-8501, Japan

E-mail: † ‡ {tohru.shimizu, satoshi.nakamura}@nict.go.jp, kawahara@i.kyoto-u.ac.jp

Abstract As automatic speech recognition and translation of long and complicated utterance cause more errors, there is increasing requirement for utterance segmentation techniques. This paper proposes speech translation unit (STU), which is a segment of an utterance which the human interpreter treats as a single cognitive unit, and also proposes speech translation unit boundary detection method by combining lexical features and prosodic features. An experimental evaluation using monologues shows that the introduction of the prosodic feature increases STU boundary detection accuracy by 7% for STU boundaries in the pause unit.

1. まえがき

自然な話し言葉(特に独話)では、良く知られているように、文の区切りが明確でない、一文が長くなる、文の途中に間(ポーズ)が空くなどの現象が見られる。これら文区切りやポーズ位置は、話者等により大きなばらつきがある。また、長い入力では音声認識や翻訳における誤りが生じやすいことが経験的に知られており、不特定話者を対象とする音声アプリケーションにおける基本的な処理単位として、文やポーズを採用するのは問題がある。また、実時間性が求められるアプリケーションにおいては、発声内容をより少ない遅延時間で処理する必要性から、入力を何らかのトリガーにより短い単位に分割することが求められている。

これまで、話し言葉を文より短い単位に分割して処理する試みとしては、主に音声から生成した字幕の改行位置を定めることを目的として、ポーズと係り受け情報に基づいて文や節にチャンキングする手法[1]、同処理の精度向上を目的として、ポーズ・形態素情報に加えて韻律情報（X-JtoBI のトーン層・BI 層のラベル）を素性として用いる手法 [2]が提案されているほか、会話の同時的な通訳を目的として、節境界やポーズの前後の形態素情報に基づいてチャンキングする手法[3]などが提案され、その有効性が示されている。

しかし、これらの処理で分割された単位（例えば、文や節）の妥当性、推定精度の目標値の設定法についての議論は必ずしも十分ではないと考えられる。また、韻律情報の利用では、利用する韻律情報ラベルが人手を介して作成されていることから、ラベルの自動推定という課題が残されている。

本稿では、分割単位の問題については、プロの同時通訳者が原言語からターゲット言語（例えば、日本語から英語）に変換する自然なタイミングである音声翻訳単位(Speech Translation Unit (STU))を提案し、音声翻訳単位の特徴を述べるとともに、音声翻訳単位境界の推定精度がポーズの有無によって大きく異なることを述べる。また、韻律情報の利用については、基本周波数(F_0)情報に基づき各形態素のF_0の平均的傾きを自動抽出し、音声翻訳単位境界の推定のための学習素性に利用する手法とその効果について述べる。

2. 音声翻訳単位コーパスの作成とその特徴

日本語の話し言葉を漸次的に英語に通訳することを想定し、日本語文の長さが長く、日本語の入力を適宜分割して英語に変換することが適当な場合、その日本語入力に対する分割位置を音声翻訳単位境界と定義した。作業は、3名のプロの同時通訳者が講演の書き起こしテキストについて行い、作業者間のばらつきを考慮し、3名中2名以上の通訳者が共通に境界と認定した箇所を音声翻訳単位境界とした。（本作業は、書き起こしテキストを用いて行っており、作業者は音声を聞いていない。書き起こしテキストは、句読点付与されておらず、200 ミリ秒以上のポーズで改行が施されているが、作業者にはポーズで改行していることを伝えていない。）

コーパス作成には、日本語話し言葉コーパス(CSJ)の 46 講演、ならびに名古屋大学同時通訳データベース[4](独話)の 16 講演を用いた。図1に音声翻訳単位の一例を示し、表1にコーパスの規模と特徴を示す。

一年間飛んでいて海外には行けるのです TP
あんまり P もっと長く飛んでいればベテランになってくれば色々行けるところもあるのです TP
いわゆる航空会社ですからただみたいなもので海外に行けるのです TP

a) CSJ

さっき入って参りましたら T
机の上に旗が立っているので TP
これは国連に来てしまったのかなと P いう感じが致しました TP
しかし T
今伊藤さんから P 別に国籍に関係なく勝手なことを言ってもいいと P いうお話がありましたので TP
安心した次第でございます TP

b) CIAIR

図1 音声翻訳単位の例
“T”は音声翻訳単位の境界位置、”P”は 200 ミリ秒を越えるポーズ位置を示す

表１ 音声翻訳単位を付与したコーパスの規模

	CSJ	CIAIR
形態素数	82,680	61,879
ポーズ(200msec 以上)がある箇所	3,844	8,826
ポーズ単位平均長(形態素数)	**21.5**	**7.0**
音声翻訳単位境界数	4,449	5,891
内訳：末尾に		
ポーズあり	3,673 (**82.6%**)	5,227 (**88.7%**)
ポーズなし	776 (17.4%)	664 (11.3%)
音声翻訳単位平均長(形態素数)	**18.6**	**10.5**

表１に示すように、本コーパスでは、CSJ と CIAIR とで、ポーズ単位の平均長、音声翻訳単位平均長の特徴に大きな違いが見られる。CIAIR では CSJ と比較して、ポーズの出現頻度が高く、音声翻訳単位あたりの平均形態素数も少ない。音声翻訳単位あたりの平均形態素数を、先行研究における区分化単位の平均形態素数で比較すると、文献[1]における節境界間の平均形態素数 9.4、文献[3]における同時翻訳単位境界あたりの平均形態素数 5.4 より長い単位であることがわかる。また、音声翻訳単位境界の多くは末尾にポーズを伴っているが、ポーズを伴わない翻訳単位境界も約１割から２割弱存在することがわかる。

次に、これまで良く用いられてきた境界である節単位と音声翻訳単位境界との比較を行った。CSJ においては、まず、日本語節境界検出プログラム CBAP[5]による節境界を求め、これを人手により修正することにより節単位としている。節単位境界のもとになった節境界と音声翻訳単位境界との比較を表２に示す。CBAP が出力する節境界は 0～3 の４種で、0 が最も強い境界である（CBAP では句点に関するルールがあることから、表２の算出時にのみ句点の挿入を行っている。その他の実験では、句点あるいは文末情報は用いられていない）。

表２に示すように、音声翻訳単位境界の多く(CSJ で 93%、CIAIR で 80%)は節境界に含まれるものの、境界が弱くなるに従って節境界であるが音声翻訳単位境界でないものも増えており、音声翻訳単位境界と節境界は異なる特徴を有していることがわかる。

表２ 音声翻訳単位境界と節境界との関係

a) CSJ

境界レベル	音声翻訳単位境界あり	音声翻訳単位境界なし
境界あり	4,154(**93.4%**)	9,823
内訳：0	3,680(82.7%)	192
1	59(1.3%)	664
2	373(8.4%)	2,936
3	42(0.9%)	6,031
境界なし	295(6.6%)	68,408
計	4,449 (100%)	78,231

b) CIAIR

境界レベル	音声翻訳単位境界あり	音声翻訳単位境界なし
境界あり	4,724(**80.2%**)	5,474
内訳：0	2,037(34.6%)	48
1	724(12.3%)	231
2	1,306(22.2%)	979
3	657(11.2%)	4,216
境界なし	1,167(19.8%)	5,0514
計	5,891 (100%)	55,988

3. プロの通訳者の音声翻訳単位境界推定精度の評価

表１のコーパス作成にあたった３名の作業者間のばらつきから、３名の作業者の多数決で決定した音声翻訳単位境界に対するそれぞれの作業者（プロの同時通訳者）の平均のＦ値を求めることができる。表３にＦ値を示す。なお、CSJ と CIAIR における作業者は異なっている。

プロの通訳者のＦ値は、CSJ と CIAIR いずれにおいても 0.9 を超えた高い値を示している。但し、単位末にポーズを伴う場合と伴わない場合を区別した場合、ポーズを伴う音声翻訳単位

表３ プロの通訳者の音声翻訳単位境界推定精度

a) CSJ

	再現率	適合率	Ｆ値
全体	93.3%	89.3%	0.912
ポーズあり	97.1%	99.4%	**0.982**
ポーズなし	76.1%	56.9%	**0.651**

b) CIAIR

	再現率	適合率	Ｆ値
全体	91.8%	88.2%	0.900
ポーズあり	93.2%	92.9%	**0.931**
ポーズなし	80.8%	60.4%	**0.691**

のF値は高く、ポーズを伴わない音声翻訳単位境界のF値はいずれも0.7を下回っており、作業者間のばらつきは、ポーズがない箇所で大きいことが読み取れる。表3に示すプロの通訳者によるF値を音声翻訳単位境界の自動推定時の上限値と考え、3節以下の実験結果との比較に用いることとした。

4. ポーズを用いた音声翻訳単位の推定

形態素情報(表層、品詞、活用形)、ポーズの有無、当該形態素が音声翻訳単位末か否かを素性とし、SVMチャンカであるYamCha[6]を用いて音声翻訳単位の学習・評価を試みた。

YamChaは以下の設定とした。

- 参照範囲：
  形態素：連続する7(前3,後3)形態素
  境界情報： 境界の前3形態素
- 多項式のカーネルの次数：2次
- 多クラスの識別： pairwise法

データ量が少ないためデータを各講演毎に分割し交叉検定を行った(CSJは46分割、CIAIRは16分割)。表3に音声翻訳単位の推定結果を示す。

表4 YamChaを用いた音声翻訳単位推定精度

a) CSJ

	再現率	適合率	F値
全体	84.3%	92.8%	0.883
ポーズあり	98.2%	98.3%	**0.986**
ポーズなし	19.7%	41.5%	0.267

b) CIAIR

	再現率	適合率	F値
全体	80.0%	82.9%	0.814
ポーズあり	86.2%	84.7%	**0.854**
ポーズなし	31.2%	55.9%	0.401

表3のプロの通訳者の推定精度と比較すると、ポーズを伴う音声翻訳単位のF値は高く、ポーズを伴わない音声翻訳単位のF値は低いという同様の傾向があることが分かる。F値の絶対値では、末尾にポーズがある音声翻訳単位では、プロの通訳者に匹敵する(CSJ)あるいは近い(CIAIR)性能が得られているものの、末尾にポーズがない音声翻訳単位のF値はかなり低い。特に再現率が著しく低いことが分かる。図2に推定された音声翻訳単位境界の例を示す。

一年間飛んでいて T
海外には行けるのです TP
あんまり P もっと長く飛んでいればベテランになってくれば色々行けるところもあるのです TP
いわゆる航空会社ですからただみたいなものでT
海外に行けるのです TP

a) CSJ

さっき入って参りましたら T
机の上に旗が立っているので TP
これは国連に来てしまったのかなと P いう感じが致しました TP
しかし今伊藤さんから P 別に国籍に関係なく勝手なことを言ってもいいと P いうお話がありましたので TP
安心した次第でございます TP

b) CIAIR

図2 推定された音声翻訳単位境界の例
“T”は音声翻訳単位の境界位置、”P”は200ミリ秒を越えるポーズ位置を示す

表4に示す結果は、学習データと評価データが同一コーパスであることから、学習データと評価データが異なる条件で、音声翻訳単位境界推定モデルの頑健性を評価した。表1、表2に示す両コーパスの特徴を勘案し、音声翻訳単位平均長が短いCIAIRで学習、CSJで評価を行った結果を表5に示す。表4と比較すると、素性としてポーズを用いた場合には、末尾にポーズがある箇所では、CSJを学習データとした場合と同程度の性能が得られたが、末尾にポーズがない箇所では、適合率が大きく低下する結果となった。CIAIRの音声翻訳単位平均長が短いことが関係していると考えられる。

表5 学習と評価のコーパスを変えた場合の音声翻訳単位推定精度
学習：CIAIR、評価：CSJ

素性	再現率	適合率	F値
全体	85.8%	74.2%	0.796
ポーズあり	98.6%	97.4%	**0.980**
ポーズなし	29.6%	16.4%	0.211

5. ポーズと F0 を用いた音声翻訳単位の推定

4節に示したように、音声翻訳単位末にポーズのない箇所の音声翻訳単位境界の推定精度の向上が課題であることから、新たな素性として F_0 から自動抽出した素性を導入する。

F_0 はフレーズ成分とアクセント成分の和として観測され、フレーズ成分の立ち上げ位置は、統語的な境界と深い関係があることがよく知られている。これまで、観測された F_0 からフレーズ成分を自動抽出する試みは行われているものの、リアルタイムでの処理を行うには計算量が多いという課題があった。そこで、観測された F_0 におけるアクセント成分の影響を軽減するために、観測された F_0 の極小値を結ぶ傾きを求め、その傾きを近似的なフレーズ成分の傾きと仮定し、この傾きを各形態素に求めたものを素性とした。以下に手順を示す。

(手順 1) 無声音部分を線形補完する
(手順 2) 手順1の結果のスムージングを行う (F_0(smoothed))
(手順 3) 手順2の結果の極小値を線形補完し (F_0(baseline))各音素の傾きを求める。
(手順 4) 各形態素の F_0 の傾き F_{0g} を構成音素の傾きの平均値として求める。

図3に F_0 情報の抽出例を示す。
また、単位末にポーズを伴う音声翻訳単位と単位末にポーズを伴わない音声翻訳単位との間の判定性能差が大きいことから、ポーズの有無によって、判定に用いる素性を変えることを考える。素性は以下の3種類とする。

- 形態素：表層+品詞+活用形
- ポーズ：ポーズの有無、ポーズ長(正規化)
- F_0：形態素の F_0 の傾き(正規化)

図4に示すように、形態素末にポーズがある場合には、形態素とポーズ情報、ポーズがない場合には、形態素と F_0 情報を用いる。

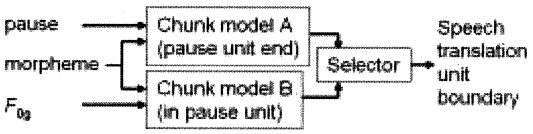

図4　ポーズの有無に基づく判定結果の統合

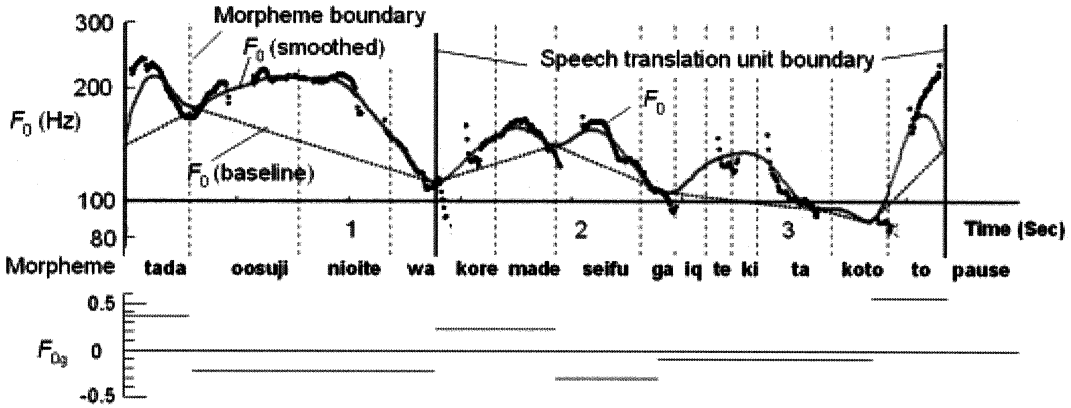

図3　F0 情報の抽出例

CSJ では音声翻訳単位付与作業時のテキストとして音声から書き起こしたテキストに修正を施したものを用いたことから、音声との対応をとることが可能なCIAIRのみについて評価を行った。表６に判定性能を示す。

表６ F_0を素性に用いた場合の音声翻訳単位推定精度（CIAIR）

形態素に加えて用いる素性	再現率(%)	適合率(%)	F値
ポーズあり			
+ポーズ	86.2	84.7	**0.854**
+ F_0	77.1	89.6	0.829
+ポーズ+ F_0	85.8	85.2	**0.855**
ポーズなし			
+ポーズ	31.2	55.9	0.401
+ **F_0**	51.4	43.0	**0.468**
+ポーズ+ F_0	29.1	59.6	0.391

表６の結果より、ポーズを伴う音声翻訳単位では、素性としてポーズ情報が有効で、素性にF_0情報を付加しても性能は変わらないことがわかる。また、ポーズを伴わない音声翻訳単位では、素性としてF_0情報を用いることによりF値が約0.07改善した。

このことから、図４に示す形態素に後続するポーズの有無で判定に用いる素性を変える処理は有効であると考えられる。しかし、プロの通訳者のF値(0.691)と比較するとF値の絶対値としては大きな改善の余地がある。

6. むすび

本稿では、日本語の話し言葉の区分化を目的として、同時通訳者の音声翻訳単位をコーパスとして整備し、同コーパスを用いて YamCha による推定実験を行った。この結果、音声翻訳単位末にポーズがある箇所では、同時通訳者に匹敵する性能が得られたが、翻訳単位末にポーズがない箇所の推定精度が低い現象が確認された。そこで、ポーズを伴わない翻訳単位の判定性能の改善を目的として、素性としてF_0から自動抽出した形態素のF_0の傾き情報を、ポーズを伴わない音声翻訳単位に選択的に用いることにより、F値の改善効果が得られることが確認された。

しかし、ポーズを伴わない翻訳単位の推定精度については、プロの通訳者による音声翻訳単位付与性能との間に大きな性能差があることから、音声情報に基づき得られる素性について引き続き検討を行う予定である。

本研究の一部は、総務省戦略的情報通信研究開発推進制度(SCOPE)(071707004)の支援により実施したものである。

参考文献

[1] 西光雅弘, 高梨克也, 河原達也, “係り受けとポーズ・フィラーの情報を用いた話し言葉の段階的チャンキング”, 電子情報通信学会技術研究報告, SP2005-137, NLC2005-104, 2005.

[2] 尾嶋憲治, 秋田祐哉, 河原達也, “局所的な係り受けと韻律の素性を用いた話し言葉の節・文境界推定”, 2007-SLP-67, pp.13-18, 2007.

[3] 笠浩一郎, 松原茂樹, 稲垣康善, “同時的な日英対話翻訳のための日本語発話文の分割”, 電子情報通信学会技術研究報告, NLC2006-56, SP2006-112, 2006.

[4] H.Tohyama, S. Matsubara, N. Kawaguchi, Y. Inagaki, “Construction and utilization of Bilingual Speech Corpus for Simultaneous Machine Interpretation Research”, Proc of 9th Europian Conf. on Speech Communication and Technology, 2005.

[5] 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝 “日本語節境界検出プログラム CBAP の開発と評価”, 自然言語処理, 11, 3, pp. 39-68, 2004.

[6] T. kudo, Y. Matsumoto, “Chunking with support vector machines”, Proc. of the 2nd meeting North American Chapter of the Association for Computational Linguistics, 2001.