

音声機械翻訳の時間効率と精度を改善するための連続音声分割

福田 りょう 須藤 克仁 中村 哲

奈良先端科学技術大学院大学

{fukuda.ryo.fo3, sudoh, s-nakamura}@is.naist.jp

概要

連続発話音声の機械翻訳には、翻訳処理単位（セグメント）への音声の事前分割が必要である。既存の音声分割手法は、文より多くの情報を持つような長いセグメントを生成する傾向があり、音声機械翻訳（Speech Translation; ST）の時間効率や精度を低下させていた。本研究では、(1) 長さに関するヒューリスティクスに頼らない分割アルゴリズムの提案、及び (2) 転移学習による音声フレーム分類器の精度改善により上記の問題に取り組んだ。TED talks の音声翻訳コーパスを用いた実験により、提案手法が生成する平均長が短いセグメントは、音声機械翻訳の時間効率と精度を改善することを示した。

1 はじめに

ST は機械翻訳（Machine Translation; MT）の一種であり、その実用化には連続音声の自動分割が欠かせない。現在の MT は適切に翻訳できる入力長に上限があるため、長い入力には事前に分割する必要がある。テキストを入力とする通常の MT では、句点や終止符を境界とした文単位への分割が一般的である。一方で、ST の入力である連続発話音声には明示的な境界記号が存在せず、セグメントの境界は自明ではない。ST モデルの学習及び評価時には、音声翻訳コーパスが含むセグメント境界を利用できるが、実用の場面では自動的な音声分割処理が必要になる。

音声区間検出（Voice Activity Detection; VAD）を用いた無音区間に基づく分割 [1] は、一般的な音声分割手法として音声認識や ST に広く用いられてきた。しかし、発話に含まれる無音は必ずしも意味的なまとまりの境界とは一致しないため、VAD はしばしば翻訳に適さないセグメントを生じさせる。近年の研究では、VAD による音声の過剰分割が ST の精度を著しく低下させることが指摘されている [2]。この問題を緩和するために、Gaido ら [3] と Inaguma

ら [4] は、ST モデルが処理しやすい長さまで VAD によるセグメントを連結して翻訳するヒューリスティクスを用いた。しかし、無音に基づく不適切な分割は依然として存在した。

近年では、音声翻訳コーパスのセグメント境界を予測する、コーパスに基づく音声分割手法 [5, 6, 7, 8] が提案されている。これらの手法は無音に基づく従来手法を大幅に上回り、中でも事前学習済みの自己教師あり音声モデル wav2vec 2.0 を用いた SHAS [7] は、人手で分割されたセグメントの翻訳精度を 95% 以上維持することが報告されている。

しかし、SHAS は 2, 3 文繋がったような長いセグメントを生成する傾向があり、より文に近い短いセグメントまで分割することで精度を改善できる可能性がある。また平均セグメント長が長いほど、並列処理できるサンプル数が減少し、かつトークンあたりの時間計算量が増加するため ST の時間効率が低下する。

そこで本研究では、翻訳に必要な情報を過不足なく持つセグメントを生成する音声分割を実現するために、SHAS に対して以下の工夫を取り入れた。

1. 長さに関するヒューリスティクスに頼らない分割アルゴリズムの提案
2. 転移学習による音声フレーム分類器の精度改善

TED talks の音声翻訳コーパス MuST-C を用いた実験により、提案手法が人手分割の約 97% の翻訳精度を維持しつつ、ST の推論速度を約 25% 高速化できることを示した。

2 従来手法 SHAS

SHAS（Supervised Hybrid Audio Segmentation）[7] は音声翻訳コーパスで学習された音声フレーム分類器（2.1 節）と確率的分割統治（probabilistic Divide-and-Conquer; pDAC）アルゴリズム（2.2 節）を用いた最新の音声分割手法である。本節では SHAS の概要と課題について説明する。

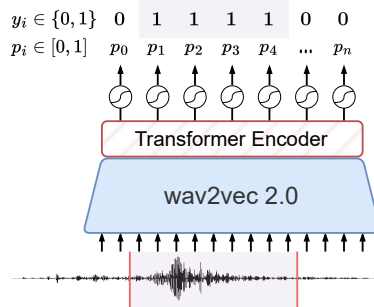


図1 SHASの音声フレーム分類器。 $y = 1$ は対応するフレームが音声翻訳コーパスのセグメントに含まれることを、 $y = 0$ はセグメント境界の一部であることを示す。

2.1 音声フレーム分類器

SHASでは、入力音声の各フレームがセグメントに含まれるか、あるいはセグメント境界の一部であることを予測する音声フレーム分類器を用いる。その構造は図1のような、事前学習済みの自己教師あり音声モデル wav2vec 2.0 のエンコーダに1層の Transformer Encoder 層を繋げたニューラルネットワークモデルである。モデルは、音声翻訳コーパスが含む高品質な分割 (Gold segmentation) を予測する、フレーム単位の系列ラベリング問題としてセグメント境界の予測を学習する。学習サンプルとして、ランダムな位置で抽出した20秒の音声と、それに対応づく20msフレーム毎のラベル $y \in \{0, 1\}$ の系列が用いられる。この時、 $y = 1$ は対応するフレームが音声翻訳コーパスのセグメントに含まれることを、 $y = 0$ はセグメント境界の一部であることを示す。学習時に wav2vec 2.0 のパラメータは固定とし、最終層の Transformer Encoder 層と出力層のパラメータのみを更新する。

2.2 pDAC アルゴリズム

音声フレーム分類器が予測した、各フレームがセグメントに含まれる確率 $p \in [0, 1]$ に基づき音声分割が行われる。pDACは、音声を最も確率が低いフレームの位置で分割し、生成されたセグメントに対しても同様の分割を適用する再帰的なアルゴリズムである。セグメントの分割は、長さが事前に設定した値を下回るか、分割の確率が閾値を上回るまで行われる。pDACの利点は、事前に設定した値を下回るとただちに分割を終了することで、STモデルが扱いやすい長さのセグメントへ音声を分割できることである。

3 提案手法

SHASは、pDACアルゴリズムで長さに関するヒューリスティクスを用いることで、安定して高い翻訳精度を達成した。一方で、SHASが生成するセグメントは音声翻訳コーパスのセグメントと比べて長くなりやすい。音声翻訳コーパスのセグメント (Gold) の平均長が5.79秒であるのに対し、実験でpDACは平均長9.17秒のセグメントを生成した(5節)。付録A.4にはセグメント長の分布を示す。以上からSHASのセグメントは、精度を低下させずに更に細かく分割できる余地があると考えられる。過大な情報は翻訳結果に悪影響を及ぼす可能性があり、更にセグメント長が長いほどSTの時間効率が低下する。本研究では、上記の問題を改善するため、長さに関するヒューリスティクスに頼らない分割アルゴリズム pTHR (3.1節) を提案する。また音声フレーム分類器の精度を向上するために wav2vec 2.0 のパラメータを一部更新する転移学習を行う (3.2節)。

3.1 pTHR アルゴリズム

pTHRは、事前に設定した確率の閾値 thr を下回る地点で分割する単純なアルゴリズムである。ただしセグメントが事前に設定した最小長 min より大きく、最大長 max 以下になることを保証する。また、隣り合う n_{ma} フレームの確率の移動平均を取ることによって予測の安定化を図っている。アルゴリズムの疑似コードを付録A.1に示す。分割時にセグメントの長さを重視していたpDACに比べ、pTHRは音声フレーム分類器の予測が結果に強く反映される。そのため過大な情報を持つセグメントが生成されにくい利点がある。また分割前に音声全体を読み込む必要があるpDACに対し、pTHRは予測に音声全体を必要としない前向きアルゴリズムであるためオンラインSTへも適用できる。

3.2 wav2vec 2.0 の転移学習

pTHRは結果が音声フレーム分類器の精度に左右されやすく、モデルの精度の改善によるSTの向上が期待できる。SHASでは、wav2vec 2.0 のパラメータを固定して音声フレーム分類器を学習した。本研究では、wav2vec 2.0 の高い表現力を更に活用するために、一部パラメータの更新を行なうSHAS+FTPTを提案する(図2)。具体的には、

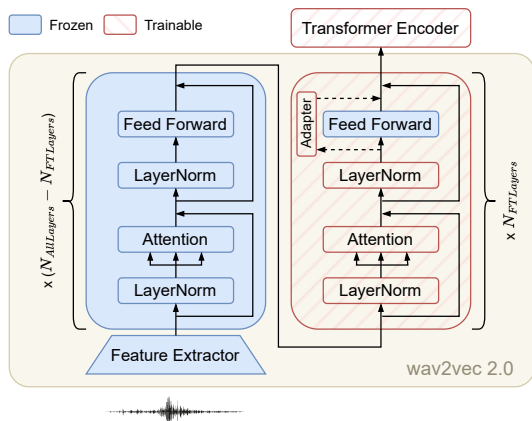


図2 wav2vec 2.0 の部分的な転移学習を行う SHAS+FTPT

wav2vec 2.0 から引き継いだエンコーダ $N_{AllLayers}$ 層の中で上 $N_{FTLayers}$ 層のパラメータを学習する。またメモリ効率を高めるため、Feed Forward 層のパラメータは常に固定として Parallel Adapter [9] を追加・学習した [10]。

4 実験設定

提案手法の有効性を検証するために音声翻訳の実験を行い、複数の音声分割手法を比較した。

4.1 データ

音声翻訳コーパス MuST-C に含まれる、約 408 時間の英語の講演音声と、それに対応付いた書き起こしテキスト、及びドイツ語の翻訳テキストを用いて英独音声翻訳の実験を行った。評価のために、学習データを用いて作成した音声フレーム分類器 (4.3.1 節) と分割アルゴリズム (4.3.2 節) で評価データの音声をセグメント化した。セグメント化された音声を ST モデルで翻訳した後、編集距離に基づくテキスト整列アルゴリズムで評価データの参照訳と対応付けを行った後、BLEU を測定した [11]。

4.2 ST モデル

ST モデルとして、Fairseq で公開されている Joint Speech-to-Text [12] モデルを用いた¹⁾。このモデルは MuST-C で学習された Transformer に基づく Encoder-Decoder 型のニューラルネットワークであり、音声とテキストの両方を入力として受け取ることができる。翻訳時、Decoder はビーム幅 5 でビーム探索を行った。

1) <https://github.com/pytorch/fairseq/blob/main/examples/speech-text-joint-to-text/docs/ende-mustc.md>

表1 BLEU による分割手法の比較。括弧内の数字は Gold segmentation との比率を示す。

Segmentation	en-de
Gold	26.99 (100.)
SHAS (Tsiamas+22)	25.67 (95.1)
提案手法	26.30 (97.4)

4.3 音声分割手法

4.3.1 音声フレーム分類器

音声フレーム分類器の学習には、学習データからランダムな位置で抽出した 20 秒の音声とそれに対応づラベル列を用いた。従来手法 SHAS の分類器として、wav2vec 2.0 XLS-R²⁾ のエンコーダを下 16 層のみ用いる *middle* モデルと、24 層全て用いる *large* モデルを作成した。また提案手法の分類器として以下に列挙する 6 つの SHAS+FTPT (3.2 節) を作成した。括弧の中に $N_{FTLayers}/N_{AllLayers}$ の形式でモデルの設定を示している。

- *middle+quarter* (4/16)
- *middle+half* (8/16)
- *middle+all* (16/16)
- *large+quarter* (6/24)
- *large+half* (12/24)
- *large+all* (24/24)

4.3.2 分割アルゴリズム

ベースラインアルゴリズムとして、pDAC (2.2 節) に加え、前向きアルゴリズム pSTRM [13] を用いた。ハイパーパラメータは $thr = 0.5$ 、 $min = 0.2$ とし、 $max = [10, 28]$ の範囲を開発データを用いて探索した中で、最良の設定で評価を行なった。提案手法である pTHR (3.1 節) のハイパーパラメータは $max = 28$ 、 $min = 0.2$ とし、 $thr = [0.1, 0.9]$ と $n_ma = [0, 1]$ の範囲を開発データを用いて探索した。このうち、移動平均を用いない $n_ma = 0$ の中で最良の設定で評価した結果を pTHR、移動平均を用いる $n_ma > 0$ の中で最良の設定で評価した結果を pTHR+MA として報告する。

5 実験結果

表 1 に SHAS と提案手法で生成したセグメントの翻訳精度を示す。ここでは SHAS は Tsiamas ら [7]

2) <https://huggingface.co/facebook/wav2vec2-xls-r-300m>

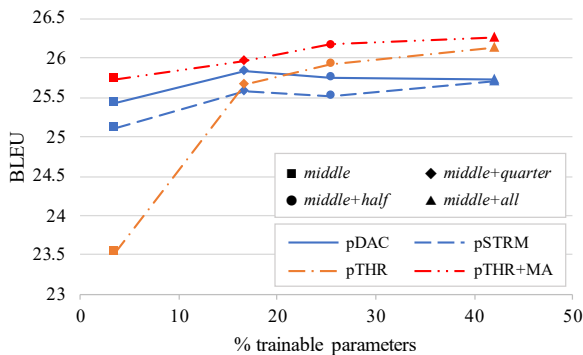


図3 4つの *middle* サイズモデルにおける分割アルゴリズムの比較。横軸はモデルの学習可能なパラメータの割合を、縦軸は翻訳精度を示す。

から引用した値を、提案手法は複数の音声フレーム分割器、分割アルゴリズム、及びハイパーパラメータで実施した中で最も高い翻訳精度を得た条件による結果を示している。より詳細な結果を付録 A.2 に示す。表 1 において、提案手法は SHAS の BLEU スコアを 0.63 pt 上回り、また音声翻訳コーパスの高品質なセグメント（Gold）の翻訳精度を 97.4% 維持した。

5.1 分類器とアルゴリズムの組み合わせ

図 3 は 4 つの *middle* サイズモデル (*middle*、*+quarter*、*+half*、*+all*) における各分割アルゴリズムの比較である (*large* モデルの結果は付録 A.3 に示す)。図の横軸はモデルのパラメータの中で学習可能なものの割合を、縦軸は翻訳精度を示している。wav2vec 2.0 のパラメータを固定して学習した *middle* では、pTHR の結果が他の分割アルゴリズムと比べて大幅に低い。この結果は分類器の予測精度が不十分であることを示唆しており、そのような場合には、長さに関するヒューリスティクスを重視した従来の分割アルゴリズムが優位性を持っていた。一方で、モデルの学習パラメータ数の増加につれ、pTHR の結果は従来の分割アルゴリズム pDAC 及び pSTRM より大きく向上した。分類器の性能が高いほど、セグメントの長さを重視する必要性が低下することが分かる。

また移動平均を用いた pTHR+MA が一貫して最も良い翻訳精度を得た。特に、*middle* では pTHR を 2pt 以上上回っており、移動平均による安定化がモデルの予測精度の低さを補う効果を示した。

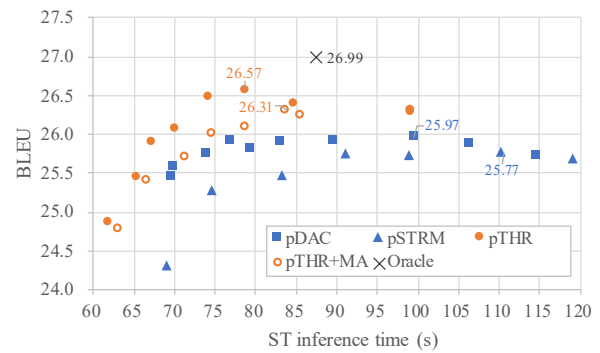


図4 分割アルゴリズムによる ST の時間効率と翻訳精度のトレードオフ。横軸は 5 回行った ST の推論の平均実行時間を、縦軸は翻訳精度を示す。

5.2 ST の時間効率

図 4 は各分割アルゴリズムの時間効率と翻訳精度のトレードオフの比較である。ミニバッチあたりのトークン数を 100,000 に設定し、同一のマシンで RTX 3090 を用いて ST の推論を行なった。また複数の点を打つために、各アルゴリズムを異なる条件で複数回実行した。pDAC と pSTRM では $max = [2, 28]$ の範囲で、pTHR と pTHR+MA では $threshold = [0.1, 0.9]$ の範囲で条件を設定した。

提案アルゴリズム (pTHR、pTHR+MA) はベースラインアルゴリズム (pDAC、pSTRM) より優れた時間効率で、高い翻訳精度を達成した。特に pTHR が生成するセグメントは、人手分割の約 97% の翻訳精度を維持しつつ、pDAC のセグメントと比べて約 25% 高速に処理された。平均セグメント長が pDAC は 9.17 秒であるのに対して pTHR は 5.67 秒と短いことが高速化に繋がったと考えられる。付録 A.4 に、セグメント手法毎のセグメント長の分布を示す。

6 おわりに

本研究では、最新の連続音声分割手法 SHAS の問題点を示した上で、(1) 長さに関するヒューリスティクスに頼らない分割アルゴリズムの提案、及び (2) 既存手法で用いられた音声フレーム分類器の精度改善によりこの問題に取り組んだ。TED talks の音声翻訳コーパスを用いた実験により、提案手法が生成する平均長が短いセグメントは、音声機械翻訳の時間効率と精度を改善することを示した。

今後は提案手法を即時性の要求が高い同時音声機械翻訳に適用し、有効性の検証を行う。

謝辞

本研究の一部は JSPS 科研費 JP21H05054 と JP21H03500、及び JST SPRING JPMJSP2140 の助成を受けたものである。

参考文献

- [1] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. **IEEE signal processing letters**, Vol. 6, No. 1, pp. 1–3, 1999.
- [2] David Wan, Chris Kedzie, Faisal Ladhak, Elsbeth Turcan, Petra Galuščáková, Elena Zotkina, Zheng Ping Jiang, Peter Bell, and Kathleen McKeown. Segmenting subtitles for correcting asr segmentation errors. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 2842–2854, 2021.
- [3] Marco Gaido, Matteo Negri, Mauro Cettolo, and Marco Turchi. Beyond voice activity detection: Hybrid audio segmentation for direct speech translation. **CoRR**, Vol. abs/2104.11710, , 2021.
- [4] Hirofumi Inaguma, Brian Yan, Siddharth Dalmia, Pengcheng Guo, Jiatong Shi, Kevin Duh, and Shinji Watanabe. ESPnet-ST IWSLT 2021 offline speech translation system. In **Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)**, pp. 100–109, Bangkok, Thailand (online), August 2021. Association for Computational Linguistics.
- [5] Xiaolin Wang, Masao Utiyama, and Eiichiro Sumita. Online sentence segmentation for simultaneous interpretation using multi-shifted recurrent neural network. In **Proceedings of Machine Translation Summit XVII Volume 1: Research Track**, pp. 1–11, 2019.
- [6] Javier Iranzo-Sánchez, Adrià Giménez Pastor, Joan Albert Silvestre-Cerdà, Pau Baquero-Arnal, Jorge Civera Saiz, and Alfons Juan. Direct segmentation models for streaming speech translation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2599–2611, Online, November 2020. Association for Computational Linguistics.
- [7] Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. SHAS: Approaching optimal Segmentation for End-to-End Speech Translation. In **Proc. Interspeech 2022**, pp. 106–110, 2022.
- [8] Ryo Fukuda, Katsuhito Sudoh, and Satoshi Nakamura. Speech Segmentation Optimization using Segmented Bilingual Speech Corpus for End-to-end Speech Translation. In **Proc. Interspeech 2022**, pp. 121–125, 2022.
- [9] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In **International Conference on Learning Representations**, 2021.
- [10] Ioannis Tsiamas, Gerard I. Gállego, Carlos Escolano, José Fonollosa, and Marta R. Costa-jussà. Pretrained speech encoders and efficient fine-tuning methods for speech translation: UPC at IWSLT 2022. In **Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)**, pp. 265–276, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics.
- [11] Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. Evaluating machine translation output with automatic sentence segmentation. In **Proceedings of the Second International Workshop on Spoken Language Translation**, 2005.
- [12] Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. Improving speech translation by understanding and learning from the auxiliary text translation task. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 4252–4261, Online, August 2021. Association for Computational Linguistics.
- [13] Marco Gaido, Matteo Negri, Mauro Cettolo, and Marco Turchi. Beyond voice activity detection: Hybrid audio segmentation for direct speech translation. In **Proceedings of the Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)**, pp. 55–62, Trento, Italy, 12–13 November 2021. Association for Computational Linguistics.

A 付録

A.1 pTHR アルゴリズムの疑似コード

Algorithm 1 pTHR

```

1: Inputs: probs, max, min, thr, n_ma
2: Initialize:
3:   segments  $\leftarrow$  empty List
4:   start  $\leftarrow$  0
5:   thrs  $\leftarrow$  List with size max
6:   thrs[min]  $\leftarrow$  0, thrs[min :]  $\leftarrow$  thr
7:    $\triangleright$  Set threshold filter thrs
8:   probs  $\leftarrow$  MovingAverage(probs, n_ma)
9:    $\triangleright$  Apply Moving Average of n_ma frames
10: while start < probs.length do
11:   if probs[start]  $\leq$  thr then
12:     start  $\leftarrow$  start + 1
13:   else
14:     end  $\leftarrow$  min(start + max, probs.length)
15:     for i = start ... end do
16:       if probs[i]  $\leq$  thrs[i] then
17:         end  $\leftarrow$  i
18:       break
19:     append segments to Tuple(start, end)
20: return segments

```

A.2 全ての音声フレーム分類器の結果

表 2 音声フレーム分類器 SHAS、SHAS+FTPT と、分割アルゴリズム pDAC、pSTRM、及び pTHR+MA を用いて生成したセグメントの翻訳精度。

Model \ Decoding	pDAC	pSTRM	pTHR+MA
SHAS			
<i>middle</i> (0/16)	25.42	25.11	25.73
<i>large</i> (0/24)	24.41	25.18	24.78
SHAS+FTPT			
<i>middle+quarter</i> (4/16)	25.84	25.57	25.96
<i>middle+half</i> (8/16)	25.75	25.52	26.17
<i>middle+all</i> (16/16)	25.73	25.71	26.27
<i>large+quarter</i> (6/24)	25.73	25.74	26.18
<i>large+half</i> (12/24)	25.89	25.58	26.15
<i>large+all</i> (24/24)	<u>25.95</u>	25.70	<u>26.30</u>

A.3 large サイズモデルのパラメータ-BLEU 曲線

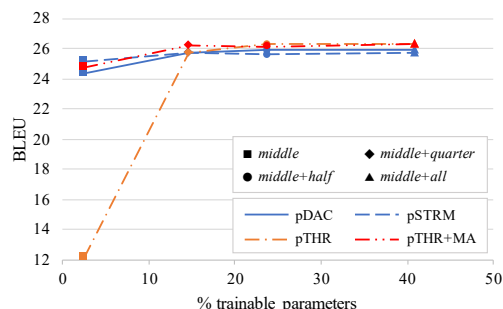
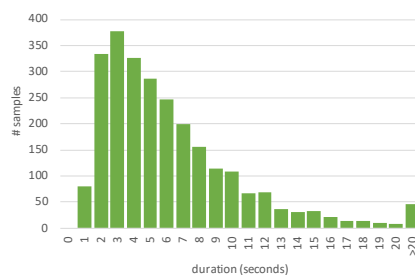
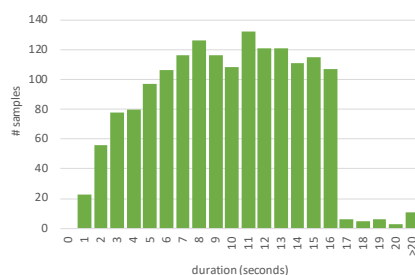


図 5 4つの large サイズモデルにおける分割アルゴリズムの比較。横軸はモデルの学習可能なパラメータの割合を、縦軸は翻訳精度を示す。

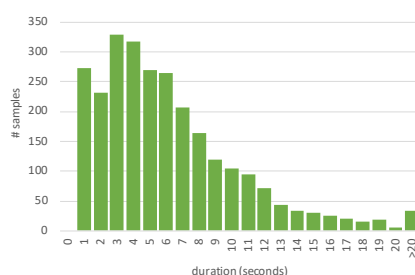
A.4 セグメント長の分布



(a) Gold セグメント



(b) pDAC セグメント



(c) pTHR セグメント

図 6 セグメント手法毎のセグメント長の分布。(b)(c)では音声フレーム分類器として *large+all* を用いた。