CPC とメルスペクトログラムの融合による音声セグメンテーション：Voice Activity Projection アーキテクチャの適応と評価

大西 一誉 [1,2,3] 池見 侑亮 [1,4] 金 道鉉 [1,2] 吉野 幸一郎 [4,3,2]
[1]mocomoco 株式会社 [2] 奈良先端科学技術大学院大学
[3] 理化学研究所 ガーディアンロボットプロジェクト [4] 東京科学大学
{kazuyo.onishi.oi5,kim.donyun.kg7,koichiro}@naist.ac.jp, ikemi.y.d23a@m.isct.ac.jp

概要

本論文では，元来ターンテーキング予測のために提案された Voice Activity Projection (VAP) アーキテクチャを，音声区間分割（セグメンテーション）タスクへと適応させる．これら二つのタスク間の本質的な関連性に着目し，高次な文脈特徴量と，低次の音響特徴量を融合するハイブリッドアプローチを提案する．Switchboard データセットを用いた実験の結果，提案モデルは CPC のみを用いた強力なベースラインを上回り，適応されたアーキテクチャの有効性と特徴量融合の重要性が示された．

1 はじめに

音声区間分割（セグメンテーション）タスク [1, 2] は，連続した音声信号から発話が含まれる区間を特定し，話者の交代点である境界を検出する技術である．音声処理における基本的な技術であり，その精度は多くの後続アプリケーションの性能に決定的な影響を与える [3, 4, 5].

近年，対話におけるターンテイキング予測モデルとして，Voice Activity Projection (VAP) [6] が大きな成功を収めている．VAP の成功の鍵は，自己教師あり学習 (SSL) モデルである Contrastive Predictive Coding (CPC) [7] の特徴量を採用した点にある．CPC は膨大なラベルなしデータから長期的な文脈依存関係を学習しており，VAP は高度な会話文脈を捉え，非常に硬い性能を達成している．

ターンテイキングを正確に予測するという VAP の目的は，本質的に「誰が話しているか，いつ話し終えるか，そしていつ次のターンが始まるか」というセグメント境界の予測タスクと同等であると考えられる．そこで本研究では，VAP のアーキテクチャは，明示的な音声セグメンテーションタスクを学習するための強力かつ合理的なフレームワークとして転用可能であるという仮説を検証する（RQ1）．

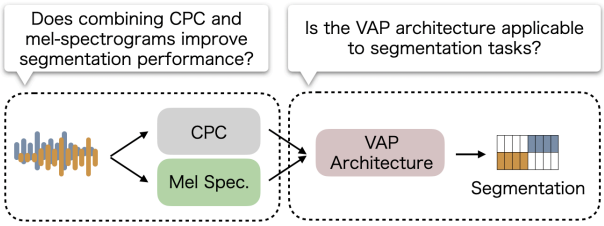


図 1: 本研究の概要

また，音声セグメンテーションタスクにおいては，近年は CPC によるマクロレベルの特徴量が用いられることが多いが，歴史的にはメルスペクトログラムのような局所的な音響的特徴が用いられてきた．CPC によって捉えられる特徴量はマクロレベルの文脈特徴であり，正確なセグメンテーションに不可欠な，瞬時のスペクトル変化などのミクロレベルの特徴をうまく捉えられていない場合がある．そこで本研究では，文脈的特徴（CPC）と音響的特徴（メルスペクトログラム）が補完的であると考え，図 1 に示すようにそれぞれの特徴量を VAP アーキテクチャで相補的に用いることで，精度向上を図る（RQ2）．

1) VAP アーキテクチャを，音声セグメンテーション専用のフレームワークとして適応・検証した．2) この特定の VAP ベースのフレームワーク内にマルチモーダル特徴量を導入し，高レベルの SSL 特徴量（CPC）と低レベルの音響特徴量（メルスペクトログラム）を融合させることで，セグメンテーション精度が大幅に向上することを示した．

表 1: 音声処理タスクの比較

用語	主な目的	出力・特徴
音声区間検出 (VAD)	発話の有無の判定	音声/非音声の 2 値フラグ
音声区間分割 (Segmentation)	境界の検出と区間特定	話者ごとの発話セグメント
話者分離 (Diarization)	誰がいつ話したかの特定	話者ラベルを含む全区間情報
音声活動予測 (VAP)	将来の音声活動の予測	短期間の将来予測（連続値）

2　音声セグメンテーション

音声セグメンテーションは，連続した音声信号から発話が含まれる区間を特定し，話者の交代点である境界を検出する技術である [3, 5]．本タスクは，後続の話者分離や自動音声認識の精度を左右する極めて重要な前処理として位置づけられている．音声処理の分野には本タスクと関連の深い用語が複数存在するため，その違いを表 1 にまとめる．

本研究が対象とする音声セグメンテーションの評価は，目的とする下流タスクにより異なるが，一般的な性能を測る指標として，検出漏れ，誤検出，話者の取り違えを統合した Segmentation Error Rate (SER) が用いられる [8]．古典的な手法では，MFCC 等の音響特徴量を用いた GMM や HMM に基づく手法が主流であった [9]．Switchboard のような自然な電話対話データセットにおける性能は，評価条件により変動するものの，一般にエラー率で 10% から 20% 程度の範囲に収まることが多い．例えば，先行研究における強力なベースラインでも，自然な対話環境下では 16.85% の SER を示している [10]．

3　関連研究

3.1　Voice Activity Projection (VAP)

VAP [6] は，現在の音声情報から各話者の短期間の将来の音声活動を連続値として予測するフレームワークであり，単純な Voice Activity Detection (VAD) [11] を超えたターンテーキングのダイナミクスをモデル化する．近年の研究では，これを表情などのマルチモーダルな手がかりやリアルタイム予測へと拡張している [12, 13, 14, 15]．

VAP のアーキテクチャは，2 人の話者からの音響特徴量をエンコードし，Self-Attention および Cross-Attention Transformer を介してコンテキスト情報を統合し，ターンテイキングを含む将来の対話的な音声活動を予測する．入力として CPC [7] などの自己教師あり学習による特徴量を用いることで，正確なターンテーキング予測に不可欠な豊かなコンテキスト理解が可能となる．

3.2　CPC とメルスペクトログラムの結合

異なるレベルの情報を捉える表現を組み合わせることは，音声処理の特徴量エンジニアリングにおける一つの傾向となっている [16, 17]．メルスペクトログラムのような低レベルの音響特徴量は，信号の局所的なスペクトル特性を忠実に表現しており，音響イベントの鋭い境界の検出に優れている．対照的に，CPC のような自己教師あり学習モデルから得られる高レベルの表現は，より広範な時間的コンテキストや言語的内容を含む，より抽象的な情報を捉える．

これら 2 種類の特徴量を組み合わせる戦略は，多くの音声言語処理タスクですでに検証されている．例えば音声変換において，自己教師あり学習による表現を用いて言語内容を制御し，同時にメルスペクトログラムを用いて音響スタイルや音色を管理する手法が提案されている [18, 19]．

4　実験

本節では，提案手法の有効性を検証するために行った実験の詳細について述べる．

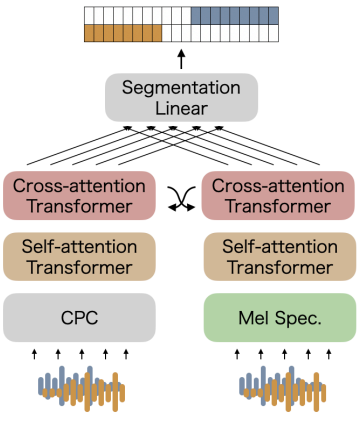


図 2: 提案モデル（VAP-Architecture / Hybrid）

4.1　比較モデル

本実験では 2 つのモデルを比較する．元の VAP は話者ごとに独立したチャンネルを想定して設計さ

れていたが，本研究ではモノラル音声を扱う．ただしマクロ，ミクロレベル双方の音響特徴のインタラクションをモデル化するため，VAP アーキテクチャの両方の入力ブランチに入力する構成をとる．

1. **ベースラインモデル（VAP-Architecture / CPC-only）：** 両方の入力ストリームに，同一のモノラル音声信号から抽出された CPC 特徴量を入力する．この構成では，並列なブランチで同一の特徴量が処理されるため，提案モデルと同一の構造でマクロ・ミクロレベル特徴量のインタラクションを考慮しない場合に相当する．
2. **提案モデル（VAP-Architecture / Hybrid）：** 図 2 に示すように，一方のストリームには CPC 特徴量を，もう一方にはメルスペクトログラムを入力する．この構成により，モデルは高レベルなコンテキスト情報と低レベルな音響の詳細を並列に処理することが可能になる．融合メカニズムには双方向のクロスアテンションを採用し，各ストリームが互いの情報を参照できるようにした．これにより，セグメンテーション性能が向上すると仮定している．

4.2 実験設定

- **データセット：** 評価には，二者間の自然な電話会話からなる Switchboard コーパス [20] を用いた．
- **入力形式：** すべての音声はサンプリング周波数 16kHz のモノラル（1 チャンネル）として処理した．モデルへの入力として，音声は 5 秒のチャンクに分割した．CPC およびメルスペクトログラムの各特徴量エンコーダは，補完なしで直接アライメントできるよう，同一のフレームレートである 20Hz（50ms ホップサイズ）で出力を生成するように設定した．
- **評価指標：** セグメンテーション性能の評価には，Segmentation Error Rate（SER）および Jaccard Index を用いた．SER は，confusion（話者誤割り当て），false alarm（誤検出），missed rates（検出漏れ）の合計であり，セグメンテーションと話者割り当ての両方の正確性を総合的に測定する指標である．
- **実装詳細：** モデルは PyTorch Lightning を用いて実装し，AdamW [21]（$\beta = (0.9, 0.999)$，weight decay=0.001，学習率=3.63×10^{-5}）を用いて最適化した．モデルパラメータは，$d_{\text{model}} = 256$，アテンションヘッド数 4，セルフアテンションレイヤー 2 層，クロスアテンションレイヤー 4 層（FFN 次元数 768）に設定した．スケジューラには ReduceLROnPlateau [22] を用い，val_loss を監視した．学習時には固定シード（1）を用い，val_loss に基づく早期終了（patience=5）を適用し，損失関数には PIT [23] を用いた BCE loss を採用した．バッチサイズは 16 とした．

表 2: 音声区間分割性能の比較

指標	CPC-only	Hybrid
Segmentation Loss	0.1957	**0.1802**
Miss Detection Rate (%)	4.04	**3.28**
False Alarm Rate (%)	3.79	**3.52**
Speaker Confusion Rate (%)	5.50	**5.40**
Total SER (%)	13.30	**12.21**
Jaccard Index (IoU)	0.8436	**0.8562**

5 結果

第 4 節で述べた 2 つのモデルについて, Switchboard のテストセットを用いて評価を行った．両モデルの評価結果を表 2 に示す．なお，評価に用いた音声の総時間は 27,387 秒である．

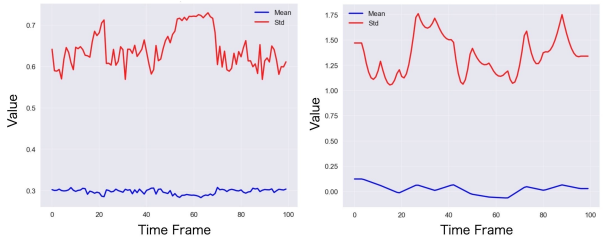
表 2 に示す通り，提案手法である VAP-Architecture / Hybrid モデルは，すべての評価指標においてベースラインの VAP-Architecture / CPC-only モデルを上回った．直接的な最適化対象である Segmentation Loss は，0.1957 から 0.1802 へと減少した．

実用的なセグメンテーション性能指標においても，顕著な改善が見られた．合計 SER は 13.33%から 12.21%へと，1.12 ポイントの絶対的な改善を達成した．SER の内訳を見ると，提案モデルは Miss Detection，False Alarm，Speaker Confusion のすべてのカテゴリにおいて誤差を削減しており，セグメンテーション品質が包括的に向上していることを示している．

6 考察

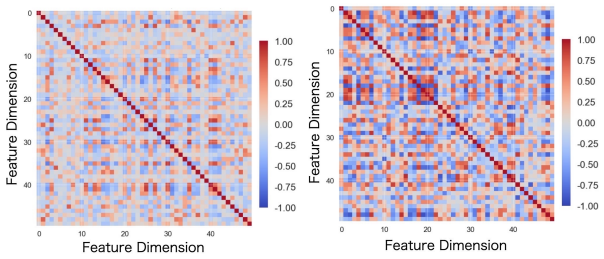
6.1 RQ1: セグメンテーションにおける VAP アーキテクチャの有効性

VAP-Architecture / CPC-only ベースラインが示した高い性能（SER 13.33%）は，先行研究のベースラインモデルの 16.85%を超える数値であり，セグメンテーションタスクにおける VAP アーキテクチャの有効性を裏付けている．本来，VAP はターンテーキ



(a) CPC (b) Mel

図 3: 時間的統計量 (Mean, Std) の比較



(a) CPC (b) Mel

図 4: 特徴量次元間の相関行列

ング予測のために設計されたものであるが，同様の問題設定であるセグメンテーションタスクに転用可能であった．また，VAP アーキテクチャに含まれる Cross-attention 機構を備えた構成においても，セグメンテーションタスクにおいて高い性能が確認された．

6.2 RQ2: ハイブリッド特徴量の優位性

本研究の主要な知見は，すべての指標において CPC-only ベースラインを上回った Hybrid モデルの優位性である．この境界検出の精度向上は，入力特徴量の統計的性質から説明できる．図 3 に示す時間的統計量の比較では，メルスペクトログラムは CPC と比較して時間方向の平均および標準偏差の変動が激しく，局所的な音響イベントの変化をより直接的に保持していることがわかる．また，図 4 の相関行列を見ると，CPC は次元間の相関が非常に高く，構造化された文脈情報を表現しているのに対し，メルスペクトログラムは各次元が独立した情報を持ち，生に近い音響エネルギー分布を反映している．

これらのデータは，両特徴量が情報量として相補的であることを示唆している．これは，自己教師あり学習による文脈的表現と，低レベル音響特徴が異なる情報粒度を捉えるという既存研究の知見とも整合的である [18]．VAP-Architecture 内のクロスアテンション機構は，CPC が提供する安定した文脈的な流れに基づき，メルスペクトログラムが持つ時間分解能の高い音響変化を適切に重み付けして統合する．このメカニズムにより，文脈情報のみでは曖昧になりやすい境界位置の特定において，物理的な音響変化を効果的に参照することが可能となり，定量的・定性的な性能向上に繋がったと考えられる．

7 おわりに

本研究では，VAP アーキテクチャの音声セグメンテーションへの適応を検討し，VAP アーキテクチャが本タスクにおいて有望な基盤であることを確認した（RQ1）．さらに，CPC とメルスペクトログラムを融合したハイブリッドモデルがベースラインを上回る性能を示し，文脈情報と音響情報の相補性が境界検出の精度向上に寄与することを実証した（RQ2）．今後は，他の自己教師あり学習モデルの統合や，話者ダイアリゼーション等のタスクへの拡張を検討する．

謝辞

本研究は，科研費 23K24910 の助成を受けて実施された．また，理化学研究所大学院生リサーチアソシエイトプログラムの一環として実施された．

参考文献

[1] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. End-to-end neural speaker diarization with self-attention. In **2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)**, pp. 296–303. IEEE, 2019.

[2] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote. audio: neural building blocks for speaker diarization. In **ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 7124–7128. IEEE, 2020.

[3] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. Speaker diarization: A review of recent research. **IEEE Transactions on audio, speech, and language processing**, Vol. 20, No. 2, pp. 356–370, 2012.

[4] Sue E Tranter and Douglas A Reynolds. An overview of automatic speaker diarization systems. **IEEE Transactions on audio, speech, and language processing**, Vol. 14, No. 5, pp. 1557–1565, 2006.

[5] Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu. End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors. **arXiv preprint arXiv:2005.09921**,

2020.

[6] Erik Ekstedt and Gabriel Skantze. Voice activity projection: Self-supervised learning of turn-taking events. **arXiv preprint arXiv:2205.09812**, 2022.

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. **arXiv preprint arXiv:1807.03748**, 2018.

[8] Alfonso Ortega, Diego Castan, Antonio Miguel, and Eduardo Lleida. The albayzin 2012 audio segmentation evaluation. **Proceedings of IberSpeech**, 2012.

[9] Lawrence Rabiner and Biing-Hwang Juang. **Fundamentals of speech recognition**. Prentice-Hall, Inc., 1993.

[10] Viet-Trung Dang, Tianyu Zhao, Sei Ueno, Hirofumi Inaguma, and Tatsuya Kawahara. End-to-end speech-to-dialog-act recognition. **arXiv preprint arXiv:2004.11419**, 2020.

[11] Koji Inoue, Bing' erJiang, Erik Ekstedt, Tatsuya Kawahara, Gabriel Skantze. Multilingual turn-taking prediction using voice activity projection. In **Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)**, pp. 11873–11883, 2024.

[12] Takeshi Saga and Catherine Pelachaud. Voice activity projection model with multimodal encoders. **arXiv preprint arXiv:2506.03980**, 2025.

[13] Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. Real-time and continuous turn-taking prediction using voice activity projection. **arXiv preprint arXiv:2401.04868**, 2024.

[14] Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation. In **Proceedings of the 11th International Conference on Human-Agent Interaction**, pp. 13–21, 2023.

[15] Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. Multimodal voice activity projection for turn-taking and effects on speaker adaptation. **IEICE Transactions on Information and Systems**, 2024.

[16] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. **arXiv preprint arXiv:2105.01051**, 2021.

[17] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. **Advances in neural information processing systems**, Vol. 33, pp. 12449–12460, 2020.

[18] Jichen Yang, Yi Zhou, and Hao Huang. Mel-s3r: Combining mel-spectrogram and self-supervised speech representation with vq-vae for any-to-any voice conversion. **Speech Communication**, Vol. 151, pp. 52–63, 2023.

[19] Jiahong Huang, Wen Xu, Yule Li, Junshi Liu, Dongpeng Ma, and Wei Xiang. Flowcpcvc: A contrastive predictive coding supervised flow framework for any-to-any voice conversion. In **Interspeech**, pp. 2558–2562, 2022.

[20] John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In **Acoustics, speech, and signal processing, ieee international conference on**, Vol. 1, pp. 517–520. IEEE Computer Society, 1992.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. **arXiv preprint arXiv:1711.05101**, 2017.

[22] PyTorch Contributors. Reducelronplateau — pytorch 2.9 documentation, 2025. Accessed: 2025-11-07 (JST).

[23] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In **2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 241–245. IEEE, 2017.