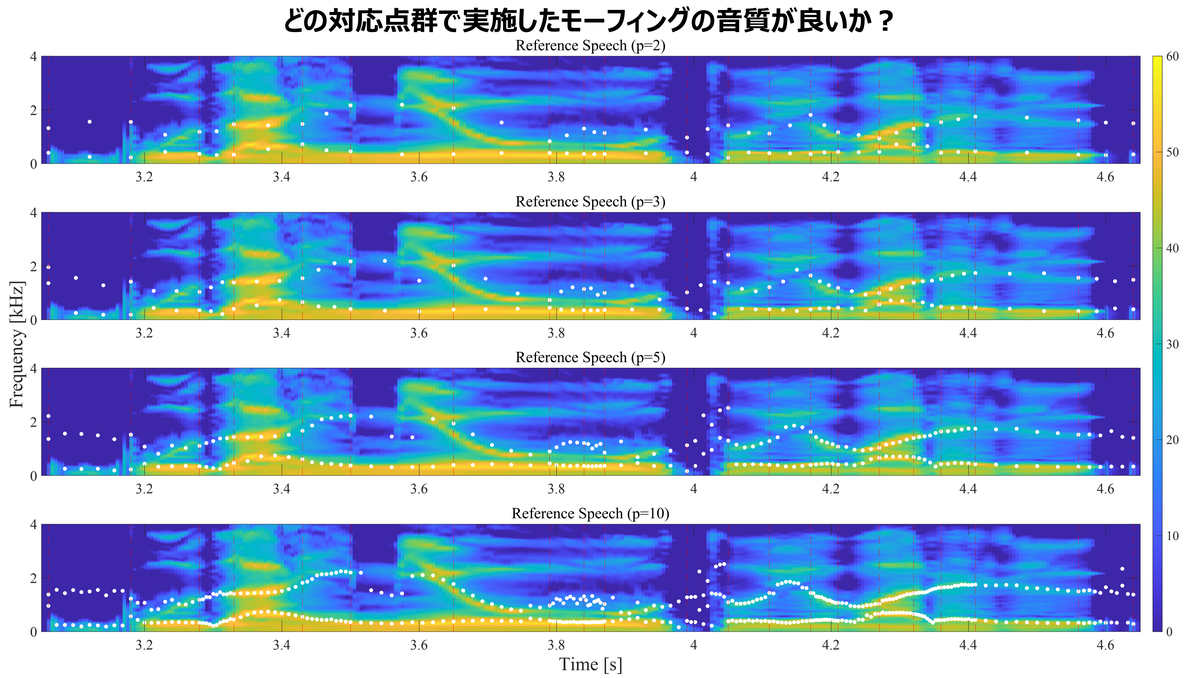
音声モーフィングにおける時間軸方向の対応点数が品質に与える影響

☆堀部貴紀[1]，森勢将雅[1]，河原英紀[2]　1: 明治大学, 2: 和歌山大学



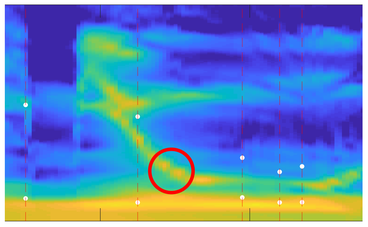
図：提案手法による対応点をプロットしたスペクトログラム
(p: 各音素区間における対応点数，丸印：対応点，赤点線：音素境界)

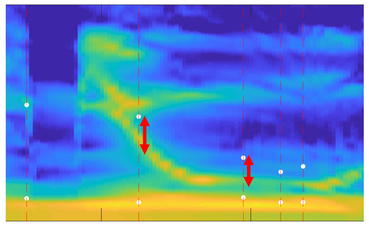
1. はじめに

音声モーフィング
同じテキストを発話する話者の2つの音声から，中間的な印象の音声を合成する技術．

対応点の自動設定手法 [堀部+ 2022]
- Juliusで推定された音素境界とLPCで得られたフォルマント周波数による対応点を検討した．
- 手動で対応付けされた音声と比較したところ，40%程度は同等以上の自然性が示された．
  → 残りは，手動によるモーフィング音声に劣る．

品質が向上する要因として考えられること
1. スペクトログラムの**変曲点**に対応点がある．
   → 時間軸方向に対応点数を増やす．
2. スペクトログラムの**軌跡上**に対応点がある．
   → 周波数軸方向の精度について検討する．


変曲点に対応点があるべき


軌跡上に対応点があるべき

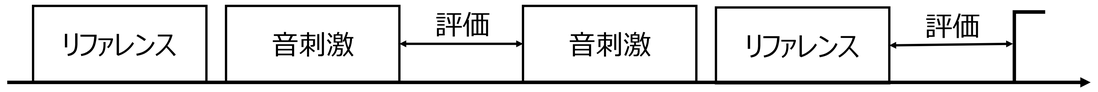
【対応点設定手法】
Juliusによる自動音素アライメント
↓
各音素区間を等分する点を求める
↓
対応点を中心に25 msフレーム切り出す
↓
前処理① 8 kHzへのダウンサンプリング
↓
前処理② 高域強調
↓
8次のLPC係数からフォルマント推定

音素境界を基準に，各音素区間における対応点数を増やすことを検討した．

2. CMOSによる主観評価実験

1-3P-7

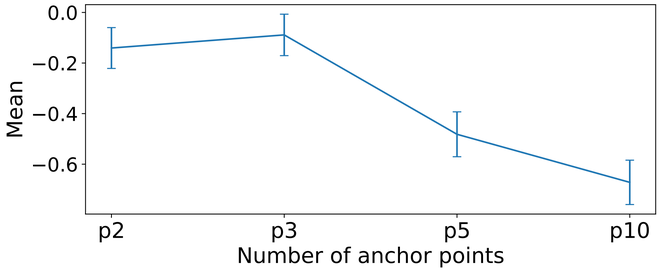
CMOS: Comparison Mean Opinion Score



- 生成されたモーフィング音声の自然性について評価する．
  - リファレンス：音素境界の対応点のみで合成されたモーフィング音声
  - 音刺激：対応点数が{2点，3点，5点，10点}で合成されたモーフィング音声
- 対応点数の違いによ る品質について評価するため，手作業によるモー フィングとは比較しない．
- 5段階評価：非常に良い[2]，良い[1]，ほぼ同じ[0]，悪い[-1]，非常に悪い[-2]

モーフィング率	50 %
使用音声DB	JVS/parallel100コーパス[1] (男女2話者ずつランダムに選択)
試行数	96試行 (4種類 × 6ペア × 2文 × 2順番) 注：ランダマイズして提示
音刺激	24 kHz，16 bits
実験参加人数	16名（正常な聴力を有する学生）

3. 結果・考察



- 縦軸：CMOSのスコア
- p2，p3，p5，p10：対応点数

- 全ての条件において，リファレンスの方が自然性が高いと判断された．
- 対応点の増加に伴い品質が低下した要因：**対応点毎のフォルマントの局所的変化**
  - （p=10における3.8~3.9秒）LPCによるフォルマントの推定結果が短時間で局所的に変化している．
    → スペクトログラムを対応点に基づき伸縮するため，品質劣化に繋がった．

2つの音声のスペクトル包絡や基本周波数は，対応点に基づいて時間軸・周波数軸において非線形に伸縮される（特に，基本周波数は時間軸方向の対応点のみ）．
→ 対応点がモーフィング音声の品質を左右する．(「音声分析合成」より)

4. 今後の展望

- 本研究では，対応点の時間軸方向に着目し，対応点数の違いによるモーフィング音声の品質について評価した．
- 主観評価では，対応点の増加に伴うフォルマントの局所的な変化で品質劣化が示唆された．
  → 時間軸方向の影響は小さい．
- 対応点の周波数軸方向について着目し，フォルマント周波数の推定手法について検討する．
  - Praat[2]を用いた音声モーフィングで品質がどのように影響するか評価する．

音源はこちらより聴くことができます．

1. S. Takamichi et al., arXiv preprint, 1908.06248, 2019.
2. B. Paul et al., Glot Int, Vol.5, pp.341–347, 2001.