



かわいらしい音声への 加工のための韻律分析

日本大学文理学部
情報システム解析学科

B4 大野 涼平



目次

- 背景
- 目的
- 従来法
- 提案法
- 分析、結果
- 音声合成、結果
- まとめ



背景

- アニメ:日本の代表的サブカルチャー
 - 世界的にも知名度が高い
 - 女性声優のアイドル化
- 「かわいらしい声」:萌え声
 - 女性がかわいらしさを大げさにした声
 - メイドカフェのメイドさん
- さとうささら, 初音ミク
 - 女子をイメージした合成音声



背景

- アニメ: 日本の代表的サブカルチャー
 - 世界的にも知名度が高い
 - 女性声優のアイドル化

• 「かわいらしい声」

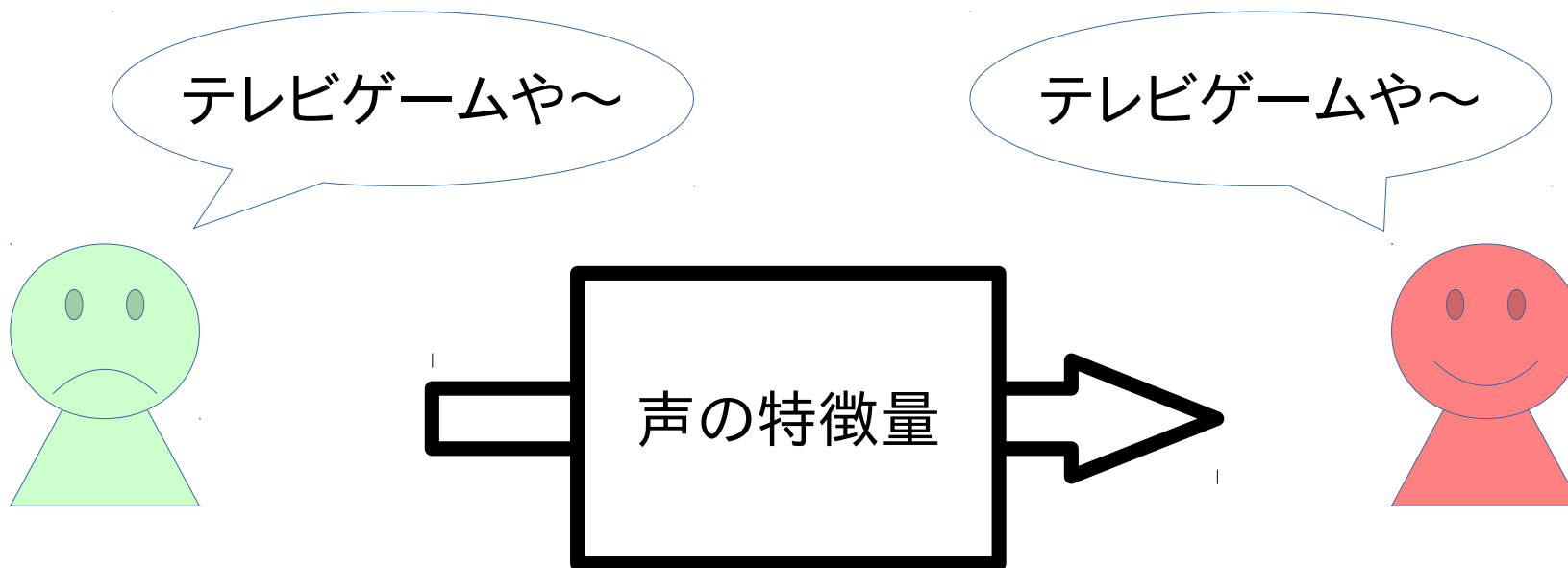
かわいらしい声に人気がある

- さとうささら, 初音ミク
 - 女子をイメージした合成音声



目的

- 「かわいらしい声」を自由に生成したい
 - 地声を任意の「かわいらしい声」に変換する



- 新たなエンターテインメントの創出に期待



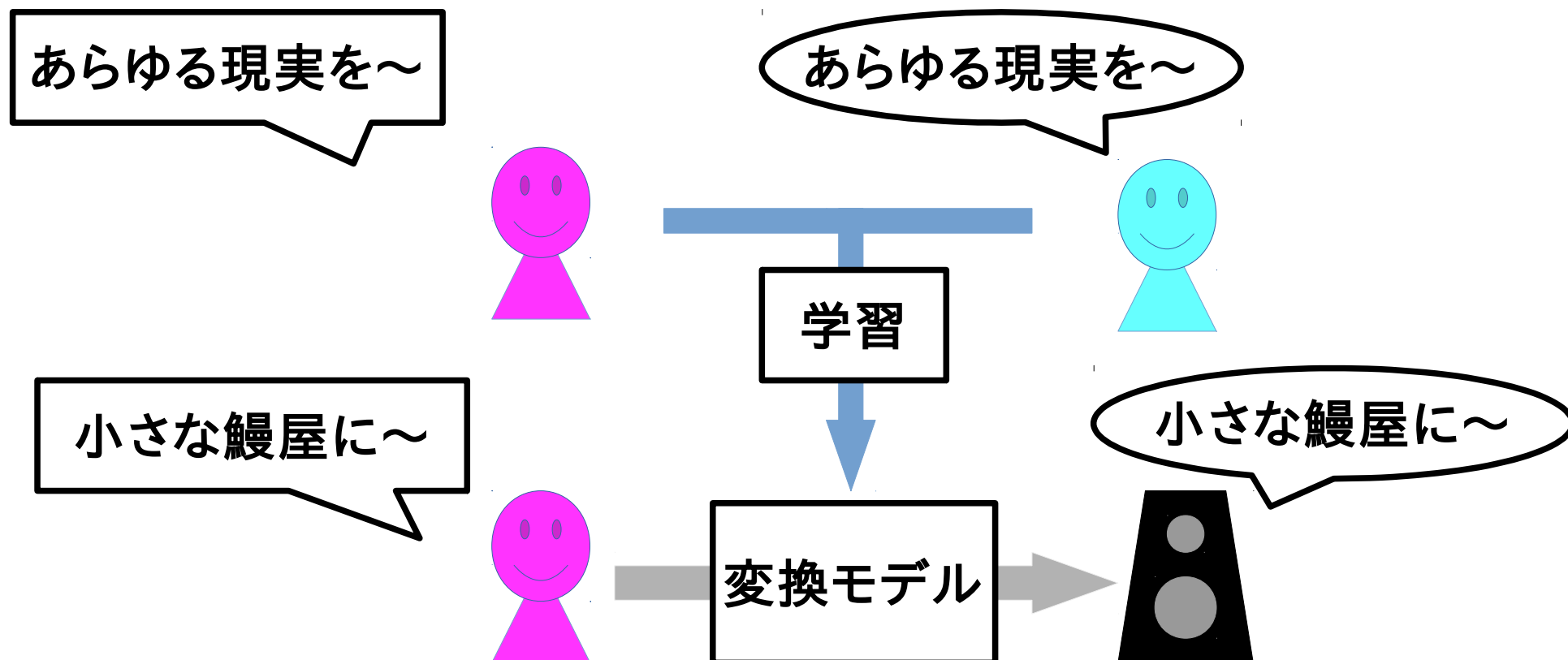
従来研究

- 声の変換
 - [1] 声質変換
 - 声の音色を変換する
 - [2] 韻律変換
 - 喋り方を変換する
- [3] 声の分析
 - かわいい音声とは？



従来研究 [1] 声質変換

- 声質変換
 - 話者の声質を他の話者の声質に変換する技術





従来研究 [1] 声質変換

- 声質変換
 - 元話者の声質を目標話者の声質に変換する技術

あらゆる現実を～

あらゆる現実を～

**声質の変換では
喋り方が転写できない**

小さな鰻屋に～

小さな鰻屋に～



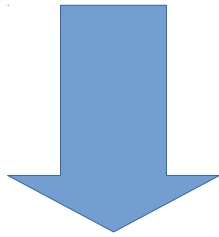
変換モデル





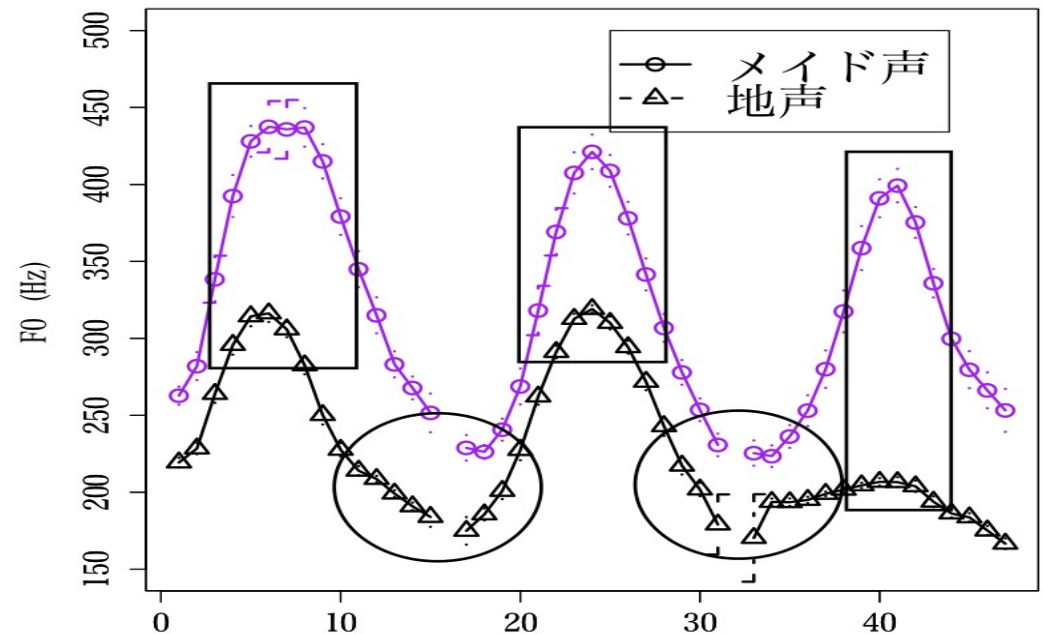
従来研究 [3] 声の分析

- 河原, 2015
 - メイドさんの地声と演技声を比較



- 演技声の方が**声が高い**
- 音高の**上昇量が大きい**
- 少し喋るのが**早い**

河原:「メイド文化と言語学」, 2015.





従来研究 [3] 声の分析

- 河原, 2015
 - メイドさんの地声と演技声を比較

河原:「メイド文化と言語学」, 2015.

セリフが限定的

- 音高の上昇量が大さい
- 少し喋るのが早い



従来研究まとめ

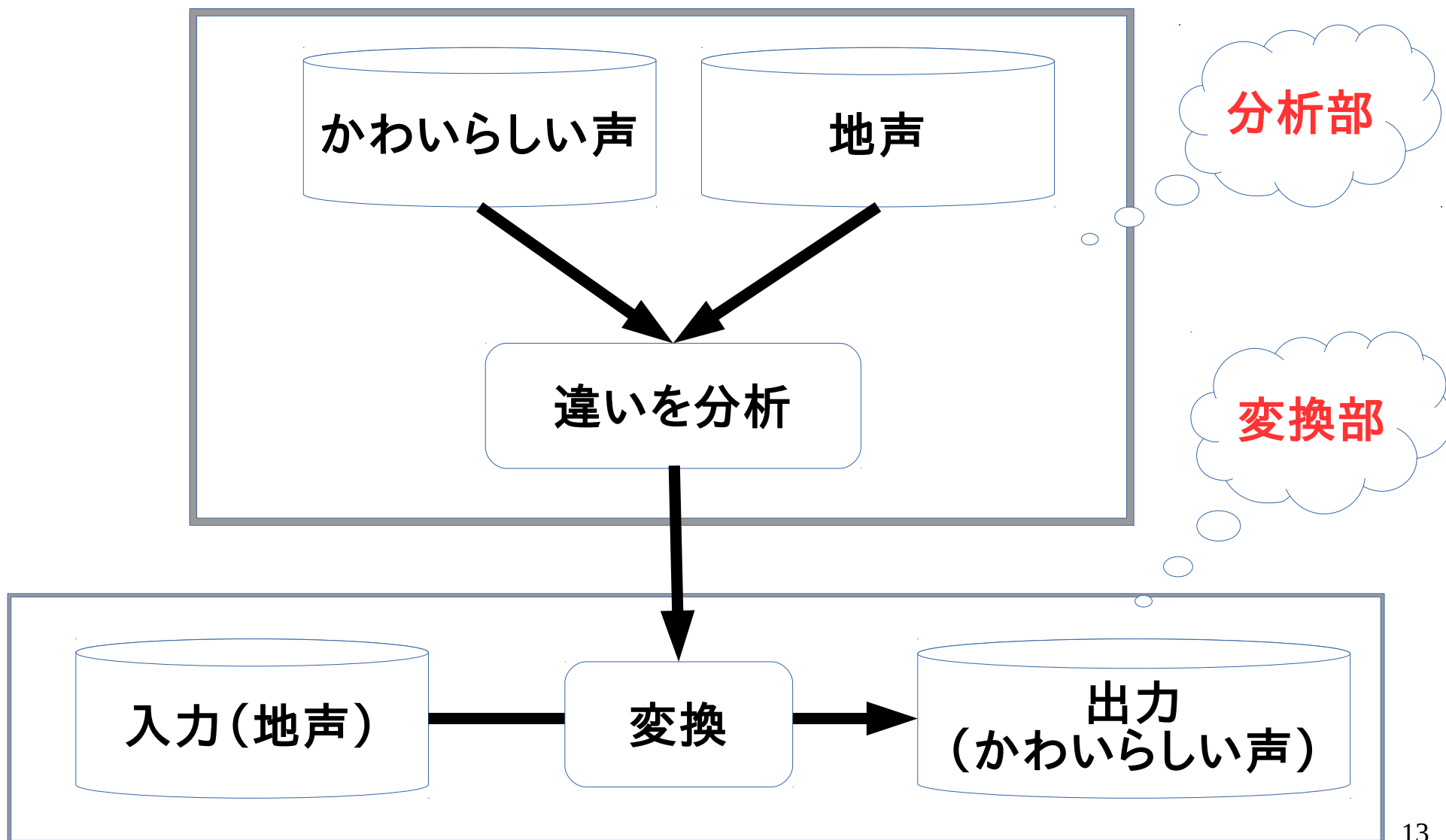
- 声質変換
 - 声優が自由に可愛さを演技できるのだから喋り方に大きな特徴がある
- 韻律変換
 - 「かわいい声」がどんな特徴かわからないから変換しようがない
- 声の分析
 - 「かわいい声」はF0が大きく関係していそう



提案法



提案法





提案法 3本の柱

- 文節単位で分析
 - かわいさは細かい変動でなく、大局的に現れる
- 文節をいくつかに分類し、F0平均値と極大値を上げる
 - かわいらしい声はF0平均値と極大値が高い
- F0を3次曲線で近似
 - よりシンプルに極大値を1つだけ表せる



提案法 3本の柱

- 文節単位で分析
 - かわいさは細かい変動じゃなく、大局的に現れる
- 文節をいくつかに分類し、F0平均値と極大値を上げる
 - かわいらしい声はF0平均値と極大値が高い
- F0を3次曲線で近似
 - よりシンプルに極大値を1つだけ表せる



提案法 3本の柱

- 文節単位で分析
 - かわいさは細かい変動じゃなく、大局的に現れる
- 文節をいくつかに分類し、F0平均値と極大値を上げる
 - かわいらしい声は**F0平均値と極大値が高い**
- F0を3次曲線で近似
 - よりシンプルに極大値を1つだけ表せる

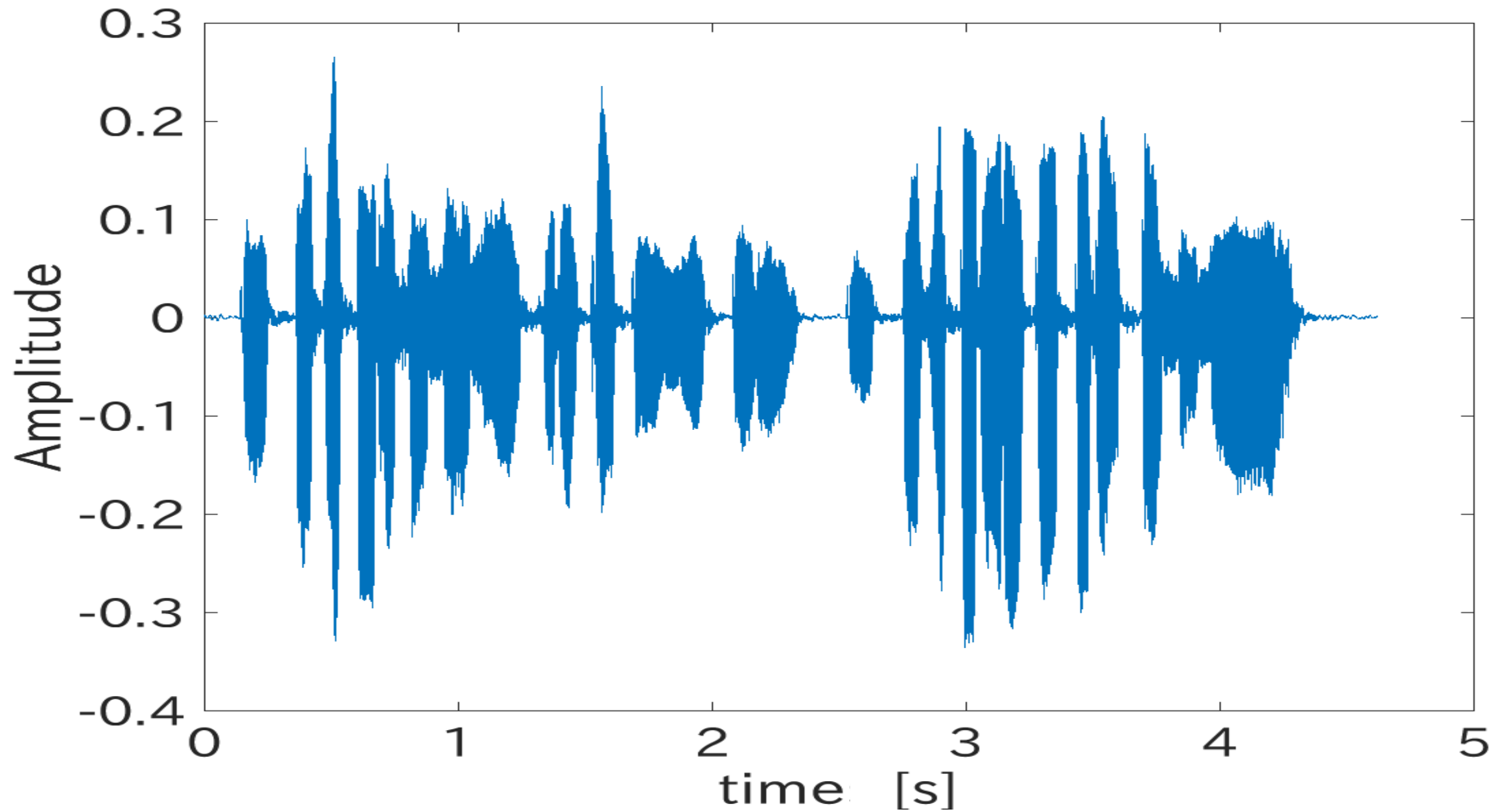


提案法 3本の柱

- 文節単位で分析
 - かわいさは細かい変動じゃなく、大局的に現れる
- 文節をいくつかに分類し、F0平均値と極大値を上げる
 - かわいらしい声はF0平均値と極大値が高い
- **F0を3次曲線で近似**
 - **よりシンプルに極大値を1つだけ表せる**



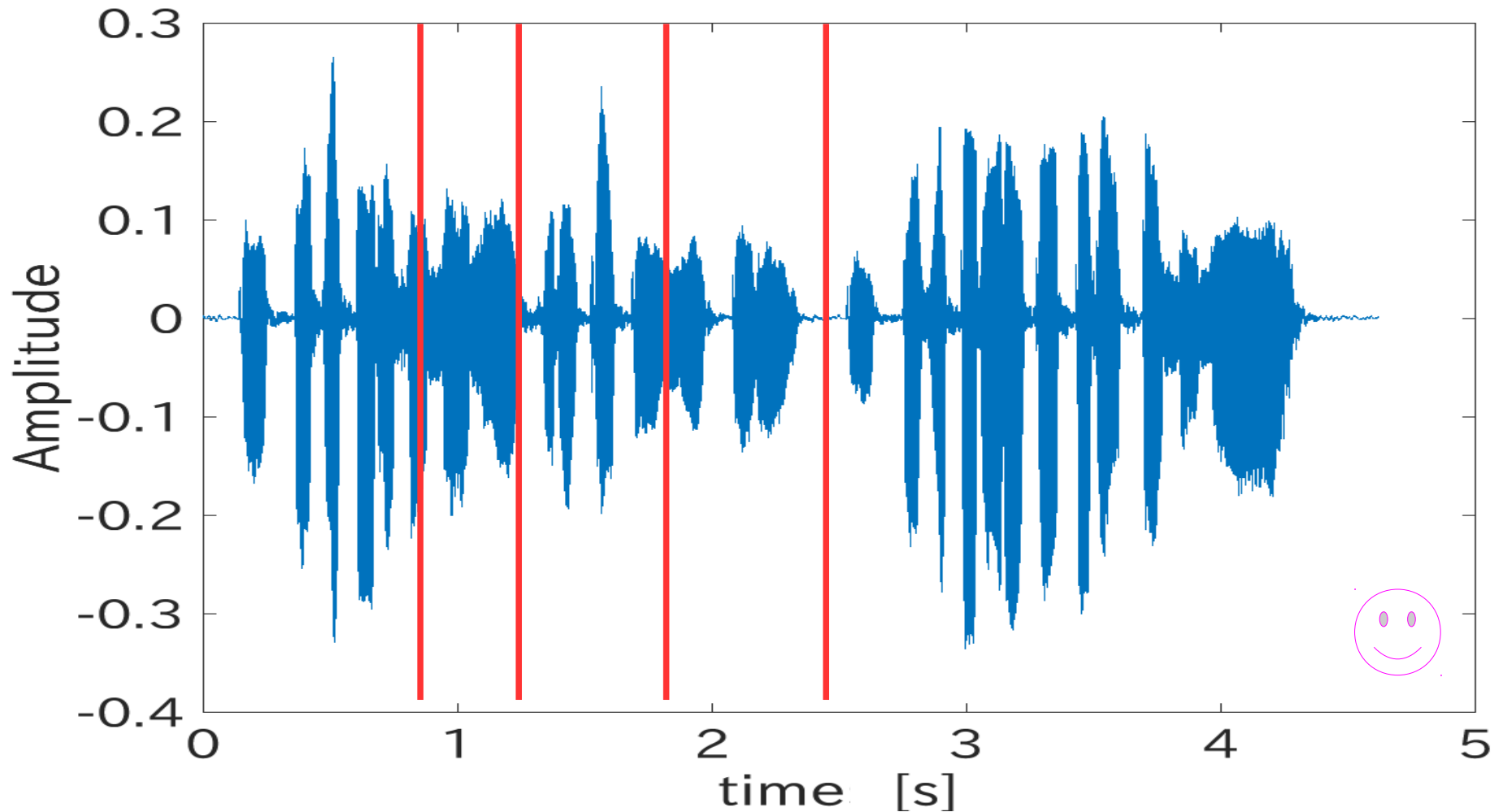
提案法 文節単位で分析





提案法 文節単位で分析

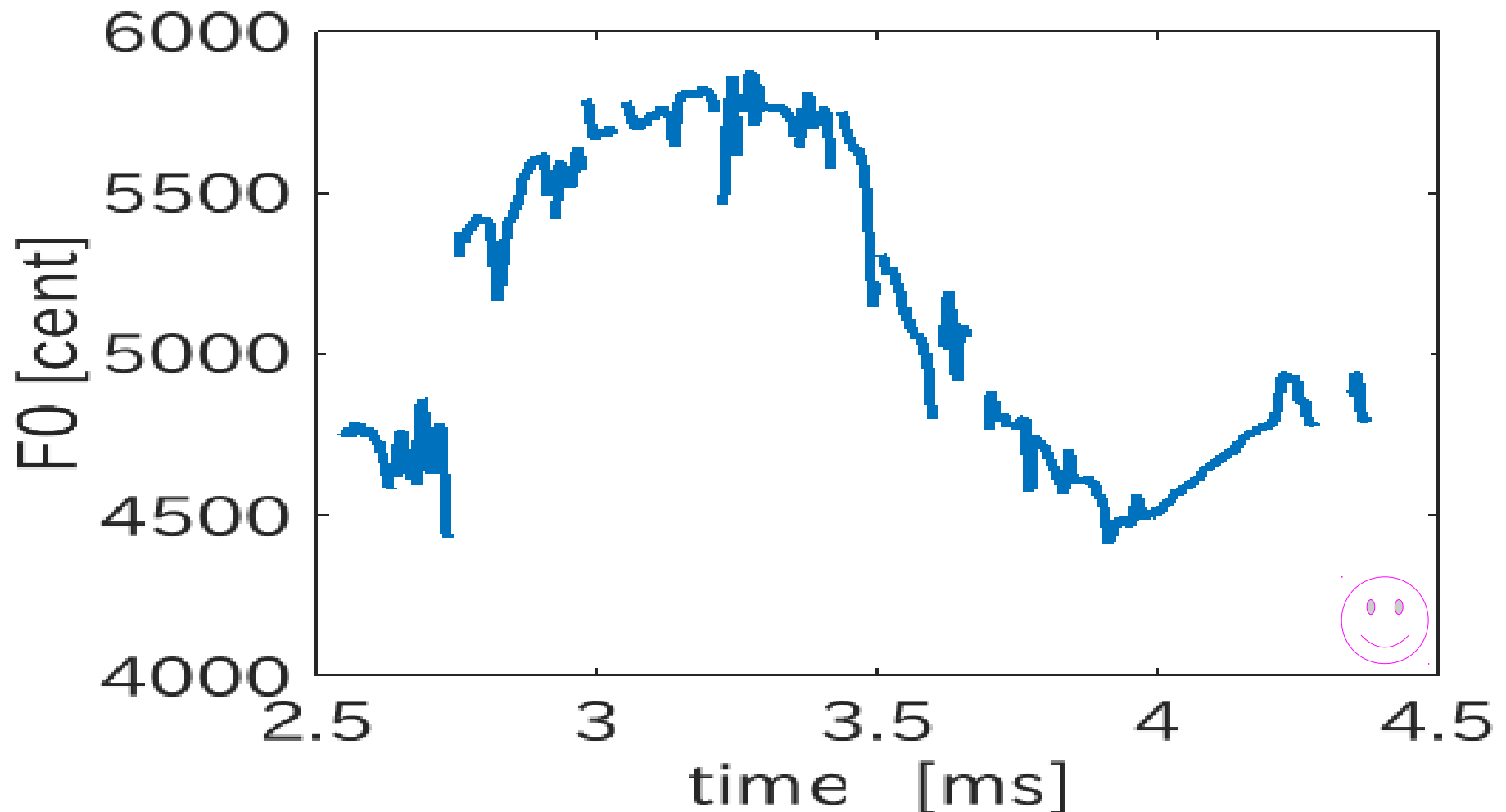
ペットボトルで するかと ペットボトルロケットだったんだねえ。
何を 思ったら、





提案法 3次曲線で近似

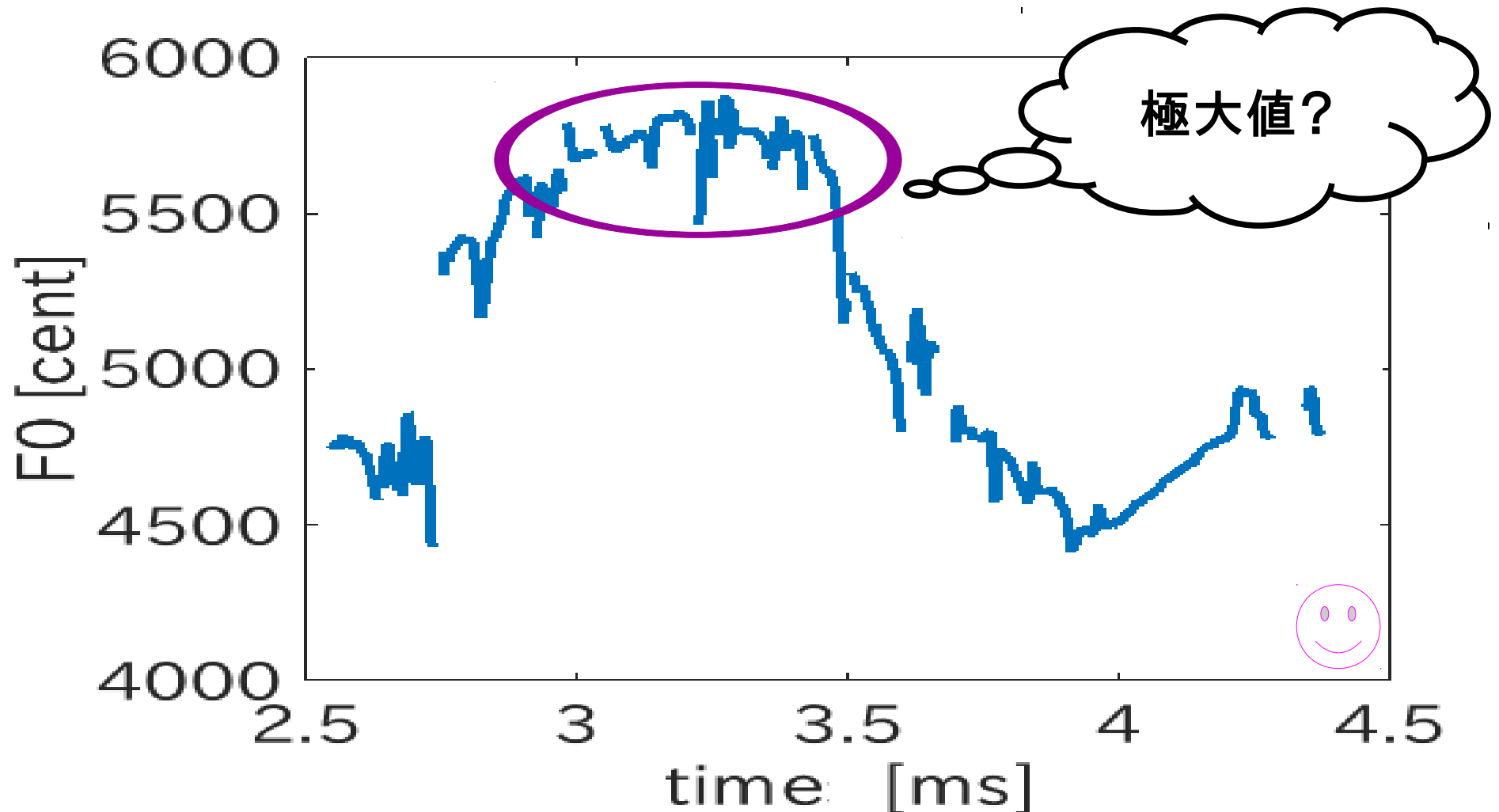
ペットボトルロケットだったんだねえ。





提案法 3次曲線で近似

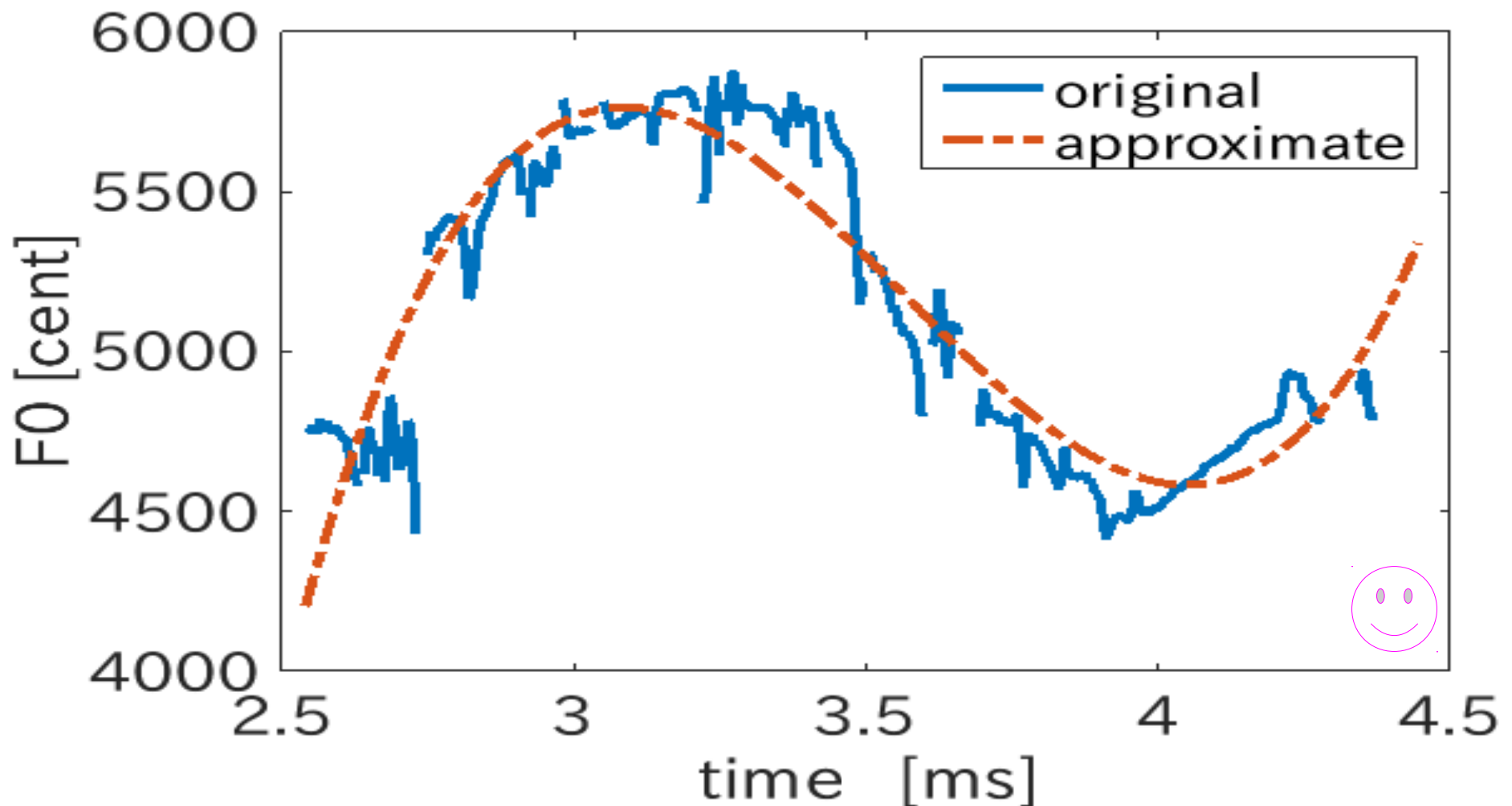
ペットボトルロケットだったんだねえ。





提案法 3次曲線で近似

ペットボトルロケットだったんだねえ。

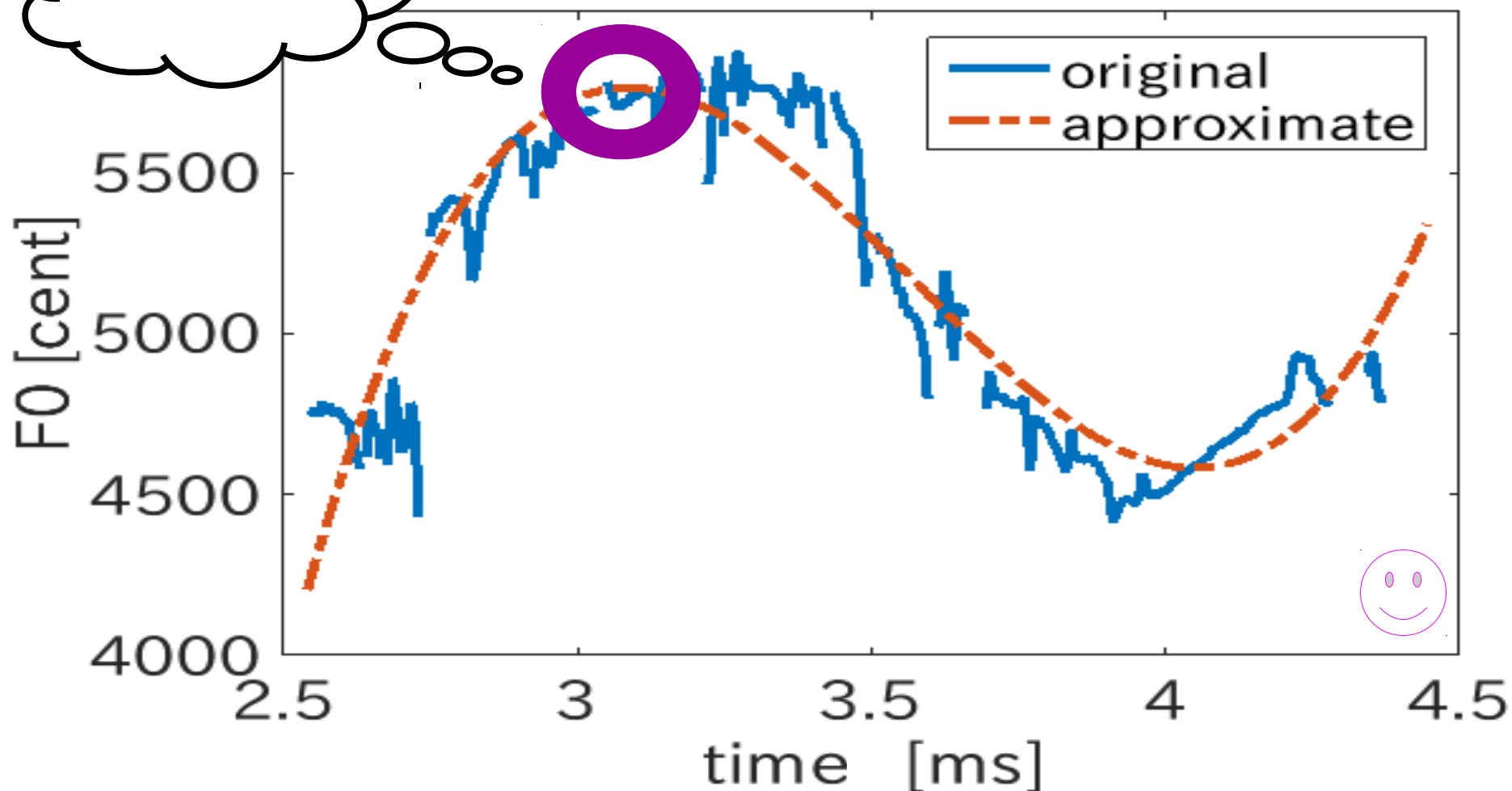




提案法 3次曲線で近似

極大値!

ペットボトルロケットだったんだねえ。





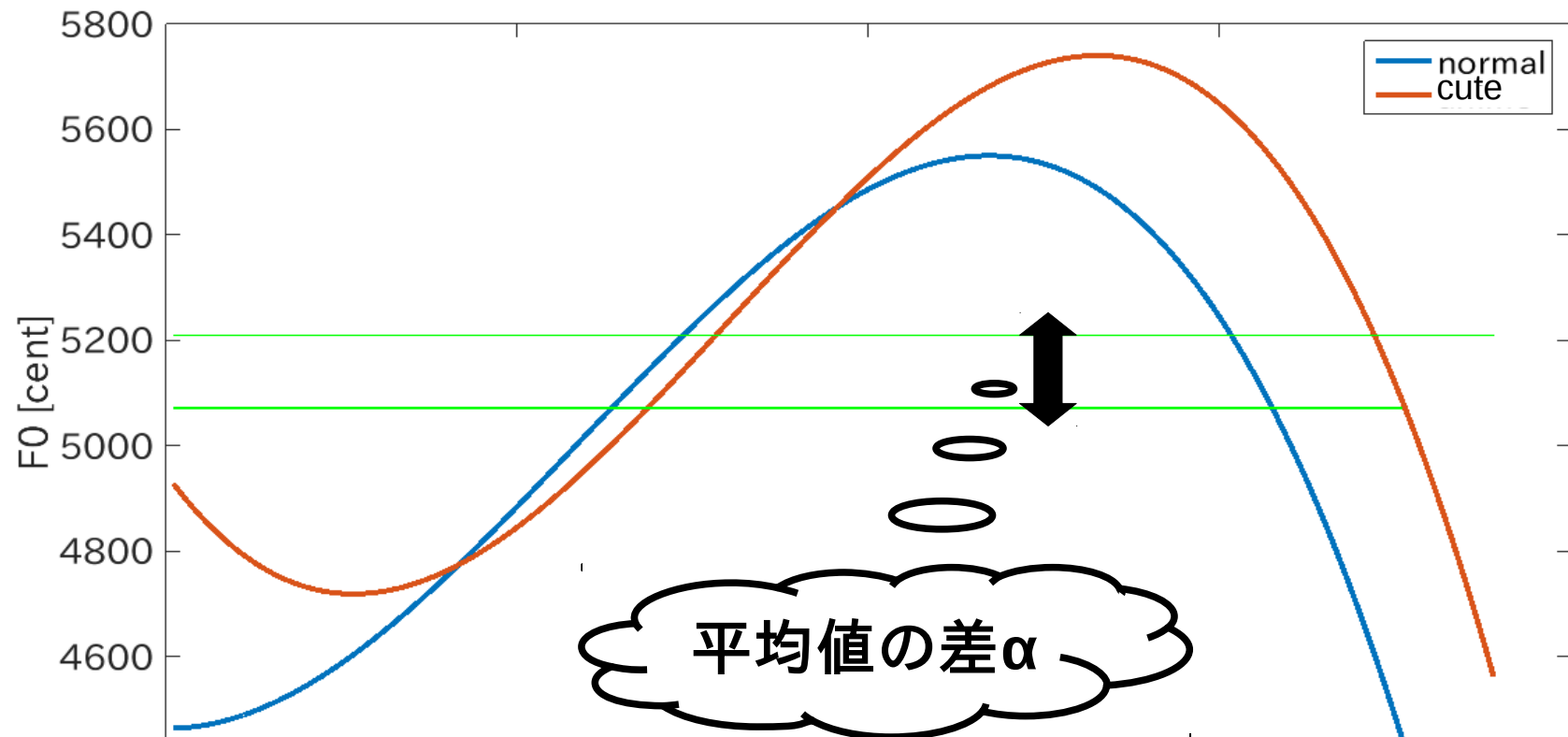
F0分析



F0分析 仮説1

- F0平均値が

▶ かわいらしい声 = 地声 + α

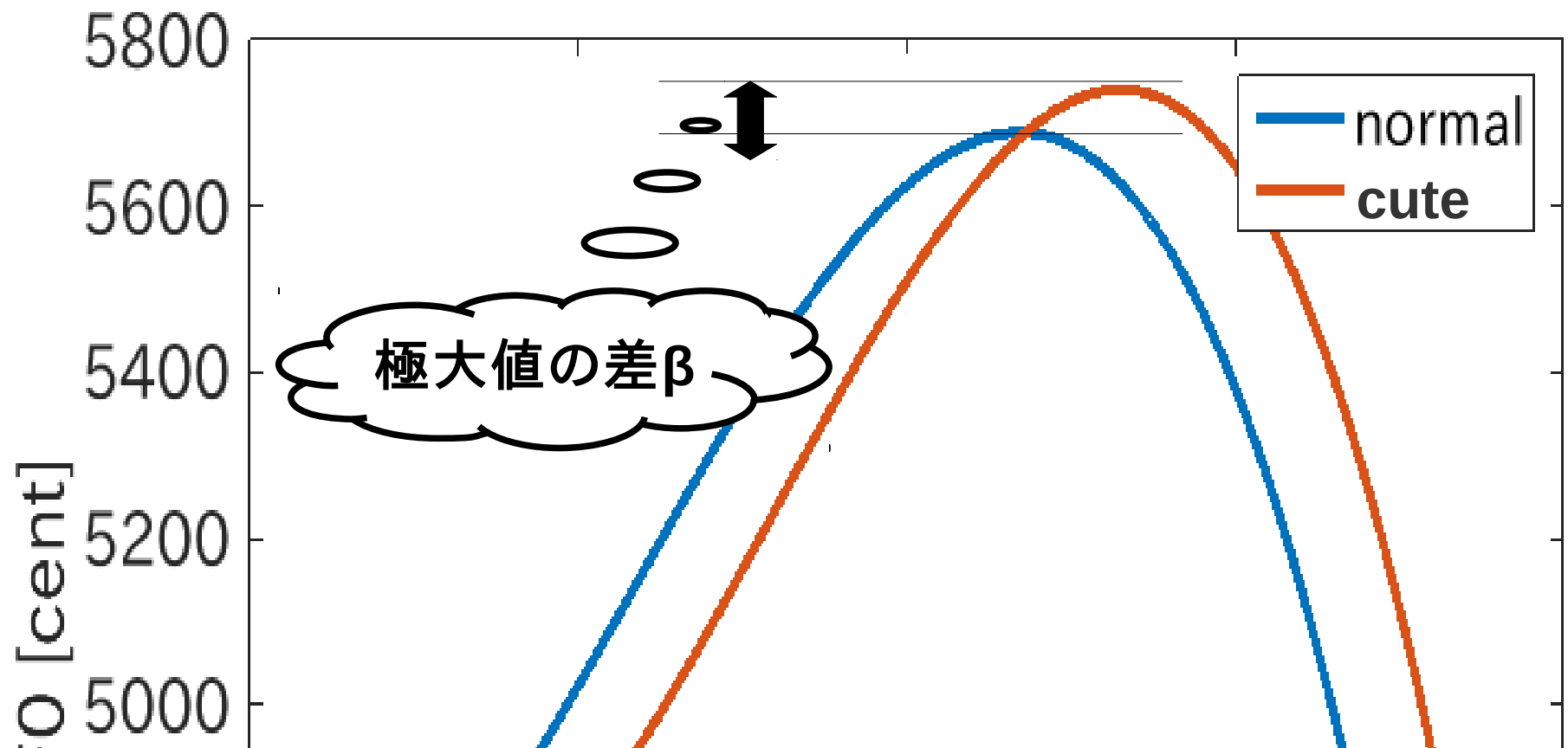




F0分析 仮説2

- F0極大値が

▶ かわいらしい声 = (地声 + α) + β





分析 使用データ

- 話者：女子大生（声優経験あり）
- 発話文：「夕映えワンダーホール」（女子高生役）
 - 地声
 - かわいらしい声（以降、演技声と呼ぶ）
- 全86文を文節に分解（全235文節）



分析 データの前処理

- F0を3次曲線で近似
- 文節にラベルをつける

文節のラベル

B	文頭(下記以外)	E	文末(句点 で終わる)
Bc	文頭(読点 で終わる)	Eq	文末(疑問 符で終わる)
S	文中(下記 以外)	Ea	文末(感嘆 符で終わる)
Sc	文中(読点 で終わる)		

- 例： 何だか、 私の 存在が ないみたいで。
(Bc) (S) (S) (E)



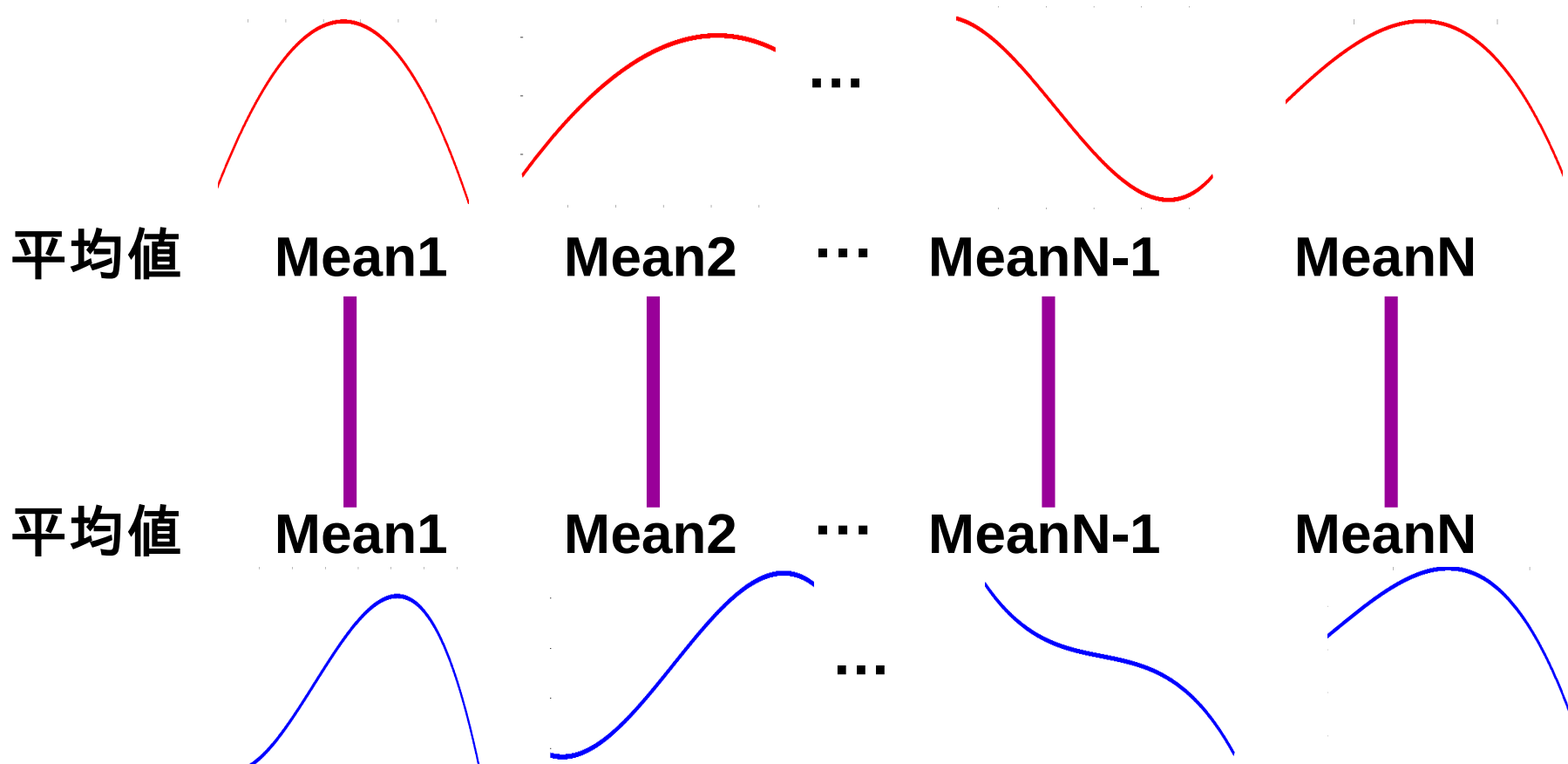
仮説1

F0平均値の分析・結果



分析 F0平均值

演技声のF0

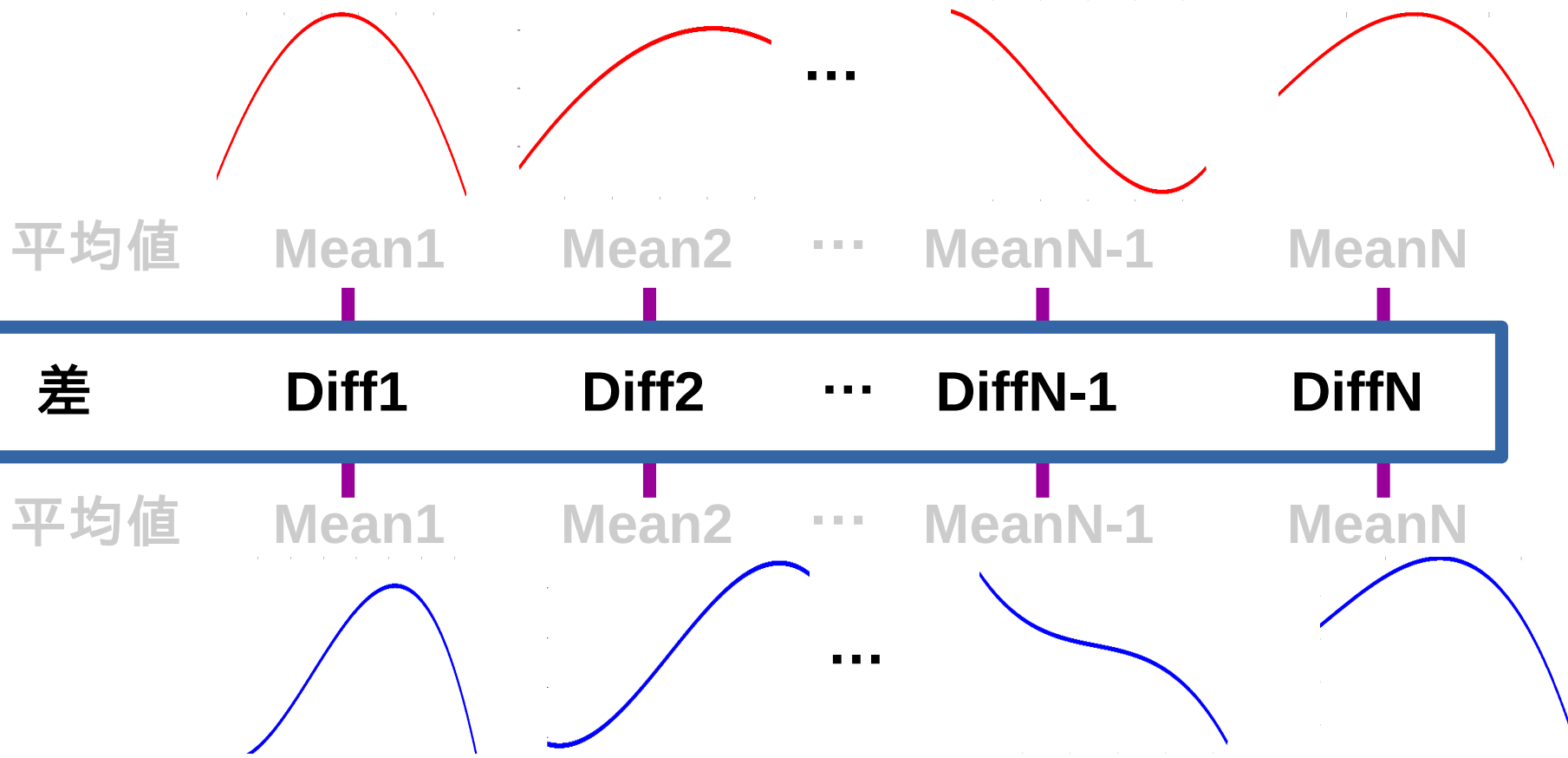


地声のF0



分析 F0平均值

演技声のF0



地声のF0



分析 F0平均値

本当に差は意味があって正になるの？
→ t 検定で調べる

差

Diff1

Diff2

...

DiffN-1

DiffN

“差”の平均値が、
2つの音声の高さの「差」



結果 F0平均値

ラベル	平均値	標準偏差	p値
B	176.7	225.3	0.0000
Bc	194.5	210.3	0.0004
S	194.4	196.2	0.0000
Sc	336.9	172.1	0.0000
E	314.9	266.7	0.0000
Eq	270.1	129.3	0.0000
Ea	211.4	170.9	0.0000
全ラベル	222.7	206.5	0.0000

- 差の平均値同士に有意な差があるか？
 - 多重比較の結果, どのラベル間の差も有意な差と言えず



仮説2

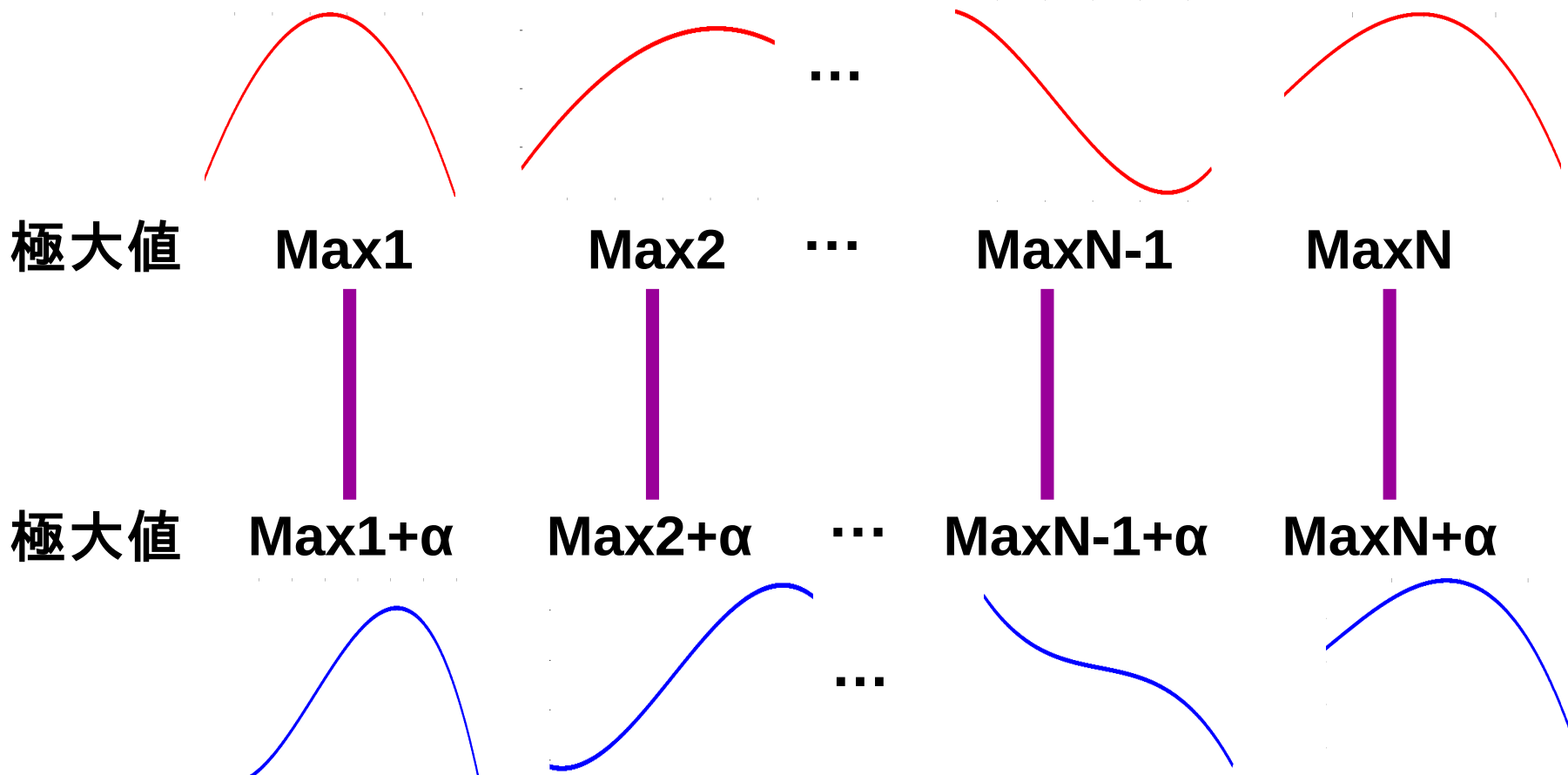
F0極大値の分析・結果



分析 F0極大値

$\alpha = 222.7$ [cent]

演技声のF0



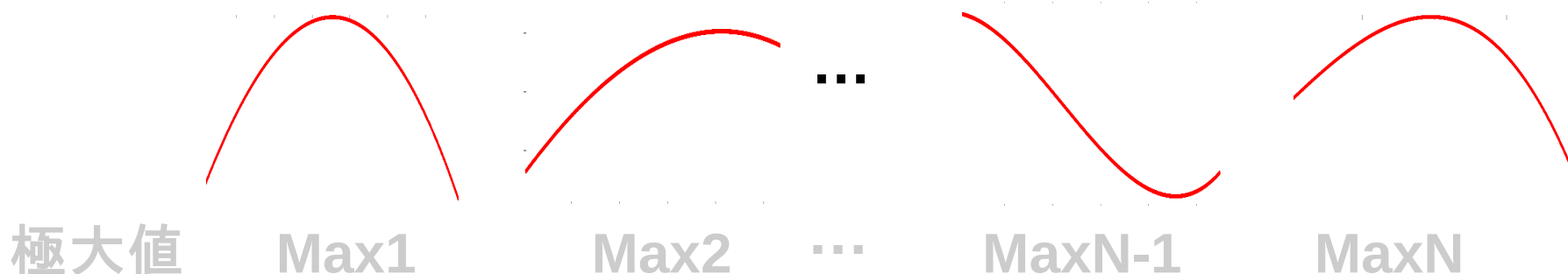
地声のF0



分析 F0極大値

$\alpha = 222.7$ [cent]

演技声のF0



差

Diff1

Diff2

...

DiffN-1

DiffN

極大値

Max1+ α

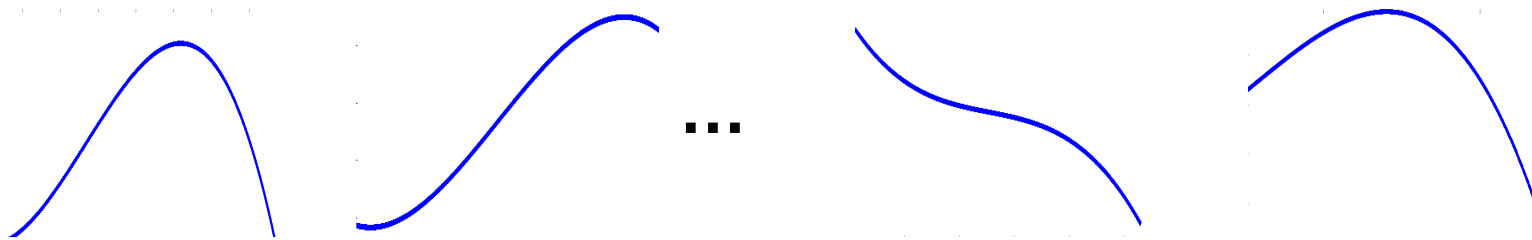
Max2+ α

...

MaxN-1+ α

MaxN+ α

地声のF0





分析 F0極大値

本当に差は意味があって正になるの？
→ t 検定で調べる

差

Diff1

Diff2

...

DiffN-1

DiffN

“差”の平均値が、
2つの音声の極大値の「差」



結果 F0極大値

ラベル	平均値	標準偏差	p値
B	-21.6	224.6	0.7049
Bc	22.2	220.2	0.3462
S	18.9	208.4	0.7597
Sc	83.8	177.3	0.0971
E	161.8	195.2	0.0004
Eq	-63.7	237.7	0.8328
Ea	-64.5	198.3	0.9377
全ラベル	139.1	190.6	0.0002

有意傾向

有意な差があると言える

- ScとEの間で, 差の平均値の差は有意な差か?
 - 検定の結果, 有意とは言えず



結果 まとめ

今回用いた音声では...

- 全体的にF0は

$$\text{演技声} = \text{地声} + 222.7 [\text{cent}]$$

- Sc, Eの極大値は

$$\begin{aligned} \text{演技声} = \\ (\text{地声} + 222.7 [\text{cent}]) + 139.1 [\text{cent}] \end{aligned}$$



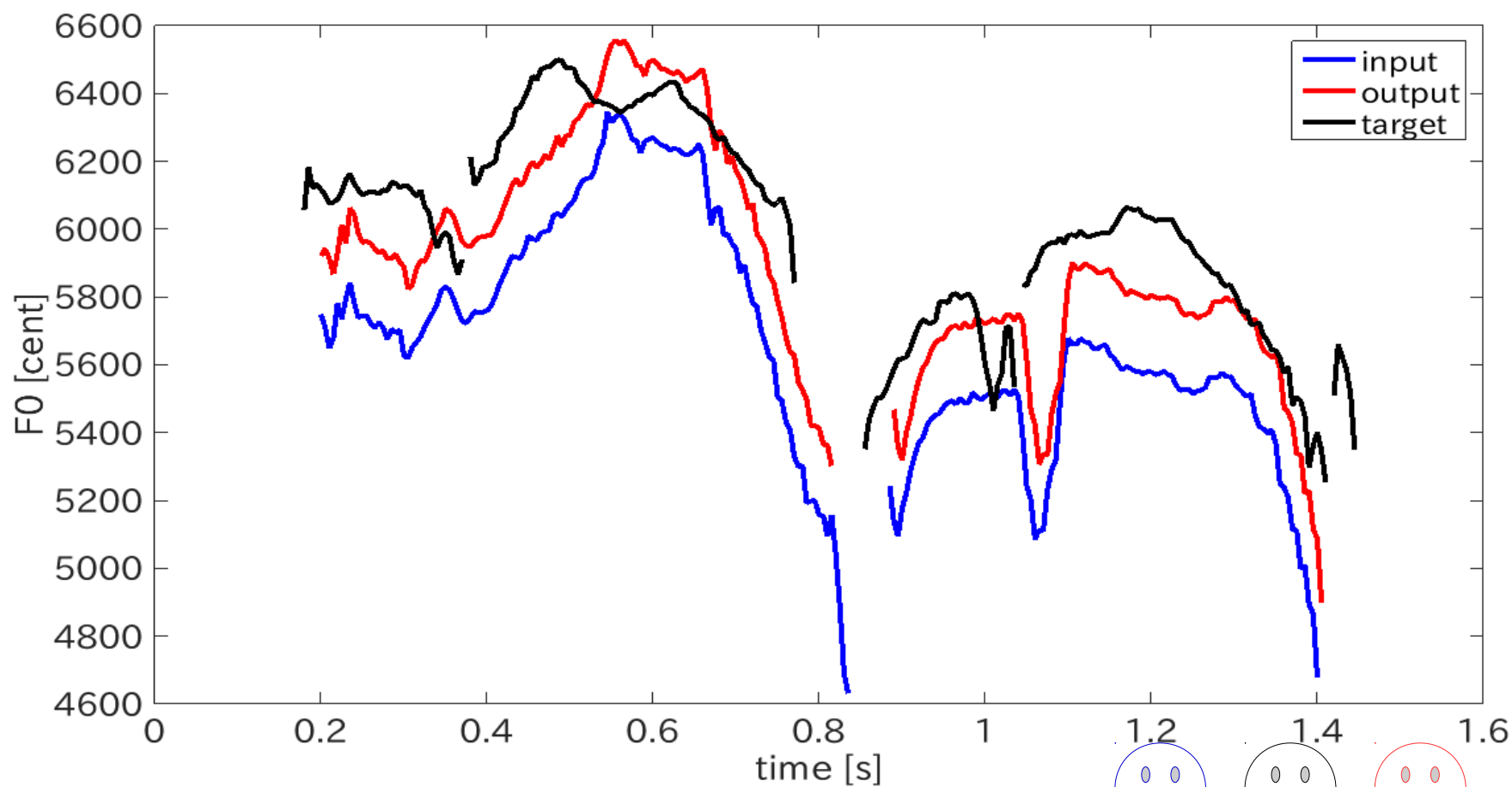
音声加工



音声加工 その1

やろうよ、(BC)

十五夜!(Ea)

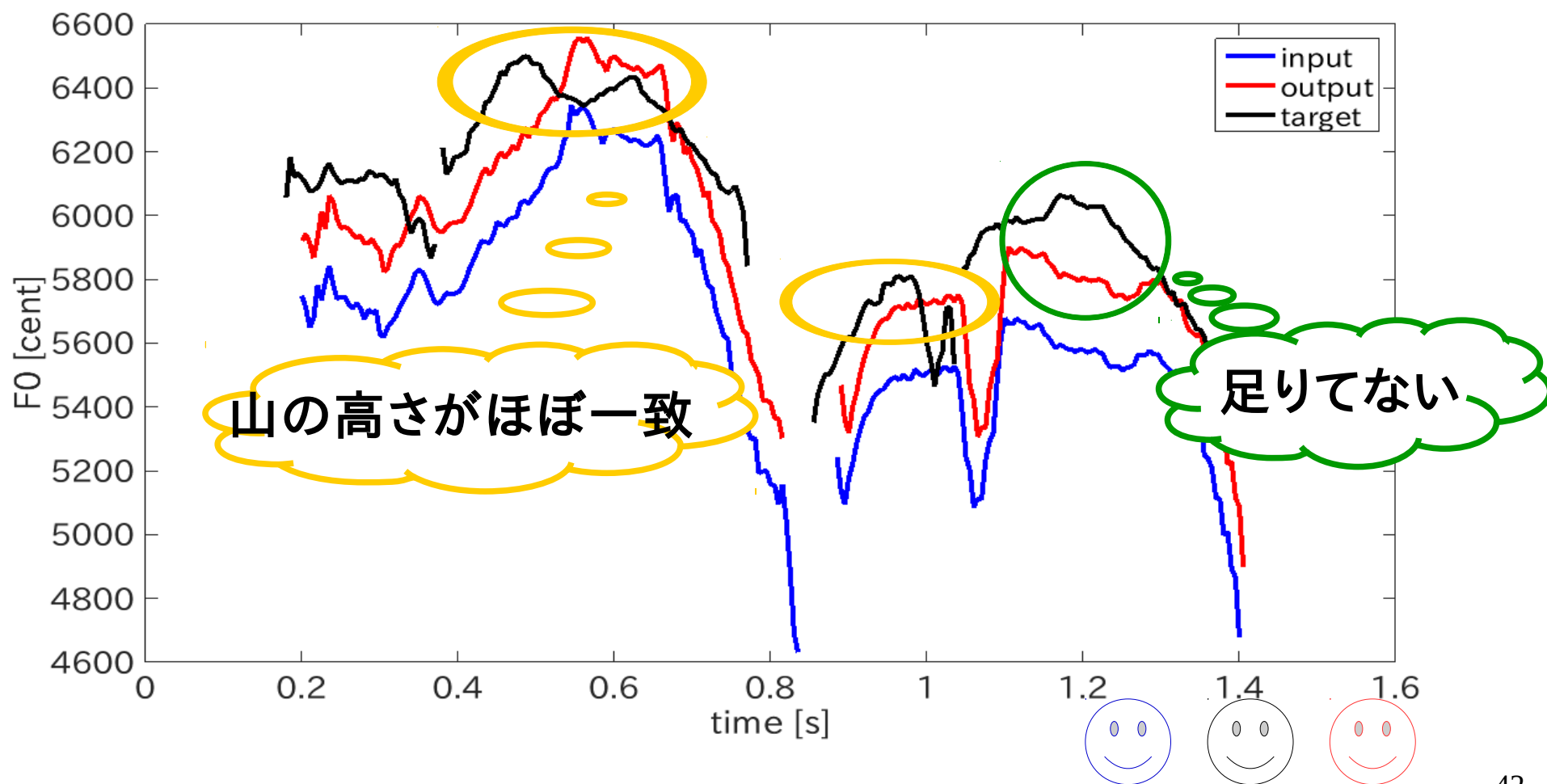




音声加工 その1

やろうよ、(BC)

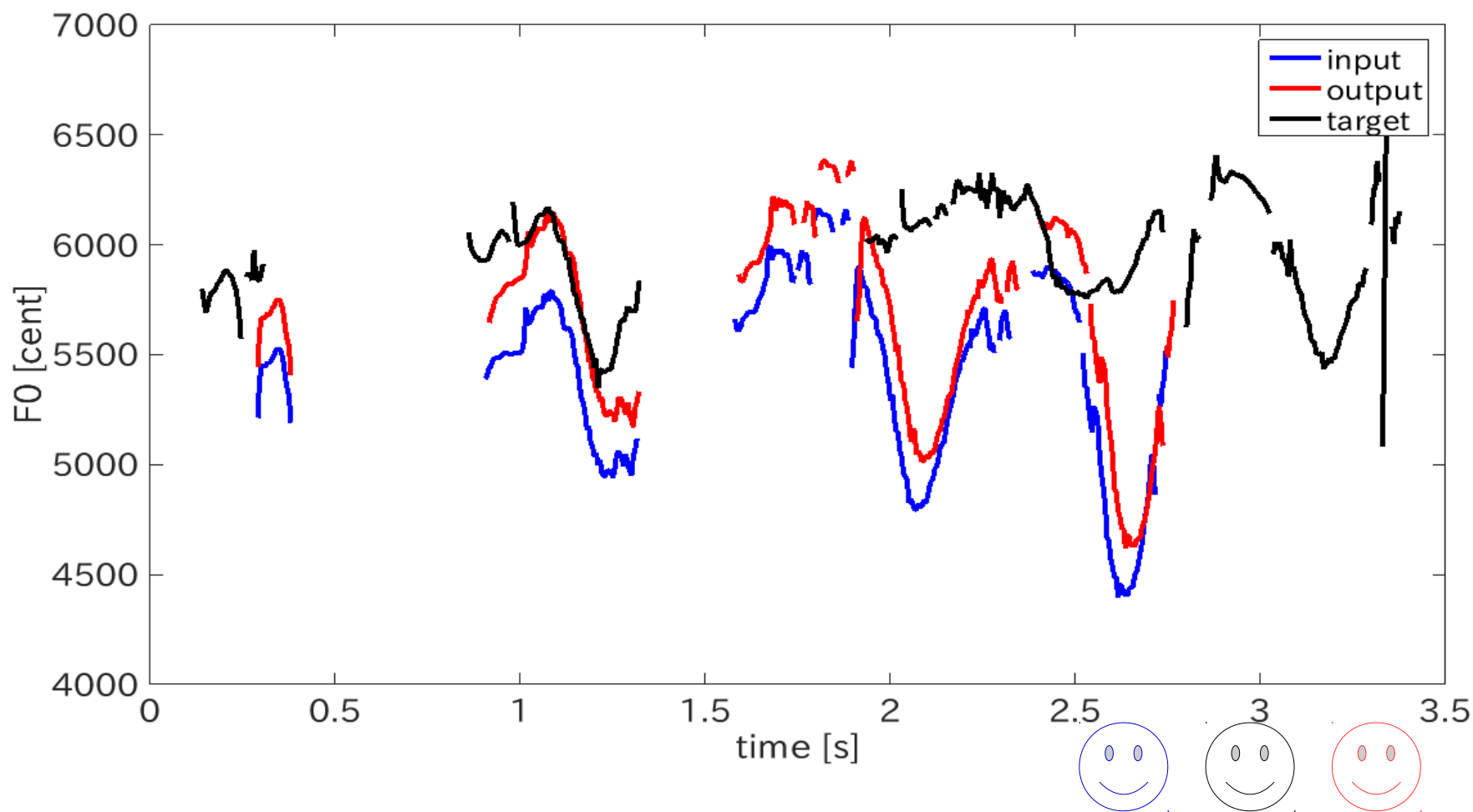
十五夜!(Ea)





音声加工 その2

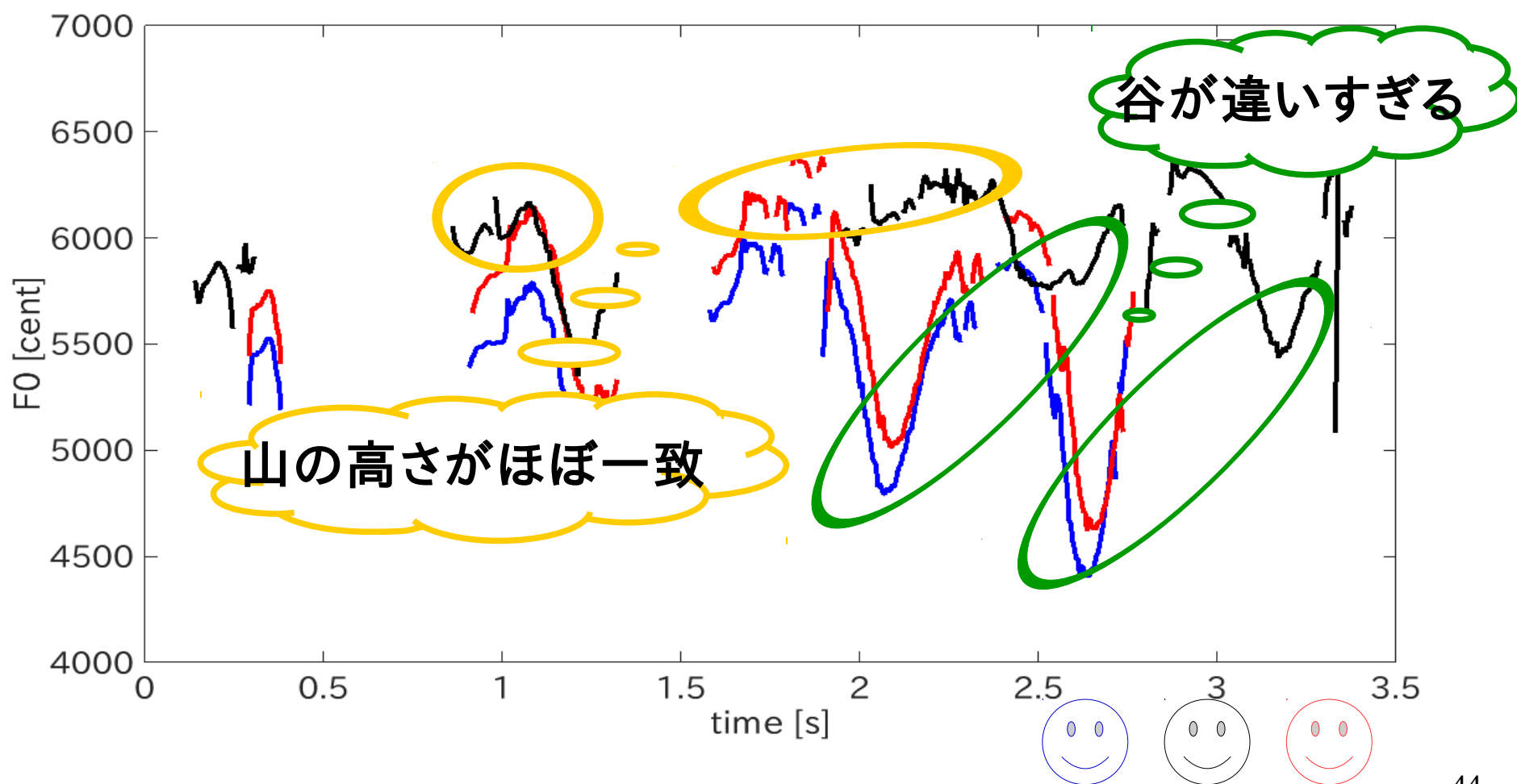
え、(BC) なに、(SC) この(S) 世界(S) 終わっちゃうの?(Eq)





音声加工 その2

え、(BC) なに、(SC) この(S) 世界(S) 終わっちゃうの?(Eq)





結果・考察

- 客観的
 - F0平均値の差はある程度埋まる
 - 極大値の増加は足りないモノが多い
 - ラベリングに問題あり
 - 平均値, 極大値以外にも考慮すべき点がありそう
- 主観的
 - 声が高くなったことは知覚できる
 - ほとんどが目標音声に近づかなかった(主観評価)



まとめ

- 地声をかわいらしい声に変換したい
 - 「かわいらしい声」とは何かを分析して変換した
- かわいらしい声は
 - 平均値 演技声 = 地声 + 222.7 [cent]
 - 極大値 演技声 = (地声 + 222.7) + 139.1 [cent]
 - (Sc, Eのとき)
- 変換結果は目標とした音声に近づかなかった
 - ラベリングに「感情」を取り入れる
 - 極大値の時間方向の位置も変換する