H-004

線スペクトル対を用いた子供と大人の音声特徴量の分布の解析

林 銀次[1,a)]　片桐 滋[1,b)]　盧 緒剛[2,c)]　大崎 美穂[1,d)]

概要：子供音声は成人音声と比べ, 声道長の違いなどから基本周波数, フォルマント周波数が大きく異なる. また, 先行研究として音声認識と音声合成を同時に可能にする音声分類器の開発があり, 音声特徴として線スペクトル対 (LSP: Line Spectral Pairs) を用いている. この特徴はフォルマント周波数周辺に密に現れることがわかっている. 以上のことから, 成人音声と子供音声の間に, LSP による分布に差が生じるという仮定のもと, 実際に分布の可視化を行い検証する. 検証の結果, 成人と子供の母音において, 比較的低い次元の LSP に分布の差を捉えることができた. また, 次元の向上と共に現れる分布の差異は, 母音によって異なる結果となった.

キーワード：音声認識, 線スペクトル対, Segmental K-means Algorithm, Code-Excited Linear Prediction

Distribution Analyses of Speech Features for Children and Adults Using Line Spectral Pairs

1. はじめに

近年の音声認識技術の普及により, 子供の音声を正確に認識する必要性が高まっている. しかし, 子供の音声は成人の音声と比べて発話内容を特定するために必要な特徴が大きく異なる. 通常, 大人の音声においては，各音素間で生じるフォルマント周波数の差を基に発話内容を分類することが一般的である. これは, 同音素間において, フォルマント周波数がある程度固有な値を持つためである. しかし, 子供の音声は成人と比べフォルマント周波数が高く, 年齢によるフォルマント周波数の変化量も大きい. これらの理由から, 成人の音声で学習をした音響モデルによって子供音声を認識すると認識率は大きく低下する [1].

先行研究として, 音声認識と音声合成の両方を同時に可能にする音声分類器の開発 [2] が行われており, 成人音声に対する正確なベイズ誤り確率の推定が可能であることが示されている. また, 子供音声に対しても, この手法の有用性が確認されている [1]. この手法では, 音声合成を行うため, 認識に用いる特徴量として線スペクトル対 (LSP: Linear Spectral Pairs) を採用している [3]. LSP は音声合成に適した特徴を持ち, さらに, フォルマント周波数周辺に密に生成される.

以上のことから, 成人音声と子供音声の間に LSP における差が生じるという仮定を基に, 本研究では Segmental K-means Algorithm (SKA)[4] によって学習した音響モデルを用いて, 各音素での LSP の差異を実験的に評価することを目的とする.

2. 特徴抽出

2.1 線スペクトル対

本論文では, 音声の特徴量として LSP を用いる. LSP は線形予測 (LP: Linear Prediction) 係数 と等価に変換可能な周波数領域のパラメータであり, 音声合成において優れた特性を持っている [5][6]. まず, 音声から特徴抽出を行う際に生じる量子化誤差による, 音声合成への影響が少ない. また, 音声合成時の時間軸方向におけるスペクトル再現精

[1] 同志社大学
Doshisha Uniersity
[2] 情報通信研究機構
National Institute of Information and Communications Technology
[a)] ctwg0106@mail4.doshisha.ac.jp
[b)] skatagir@mail.doshisha.ac.jp
[c)] xugang.lu@nict.go.jp
[d)] mohsaki@mail.doshisha.ac.jp

度が高い．さらに，音声のフォルマント周波数周辺で対となって生じる性質があり，フォルマント位置に比較的近い値を取ることがわかっている．

次に，LSP の導出過程を示す．まず，p 次の LP モデルの式を以下に示す．

$$\varepsilon_\tau = x_\tau + \sum_{i=1}^{p} a_i x_{\tau-i} \tag{1}$$

ここで，x_τ を時刻 τ における入力信号，p を LP 次数，$\{a_i\}$ を LP 係数，ε_τ を LP 誤差とする．また，式 (1) を Z 変換すると以下の式が得られる．

$$Z[\varepsilon_\tau] = \left(\sum_{i=0}^{p} a_i z^{-i}\right) X(z) = A_p(z)X(z) \tag{2}$$

ここで，$A_p(z) = \sum_{i=0}^{p} a_i z^{-i}$，$a_0 = 1$，$X(z)$ は $\{x_\tau\}$ の Z 変換であり，$A_p(z)$ は以下の式 (3)，(4) に分解できる．

$$P(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}) \tag{3}$$

$$Q(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}) \tag{4}$$

さらに，式 (3)，(4) を以下の式 (5)，(6) に因数分解することができる．

$$P(z) = (1 - z^{-1}) \prod_{i=2,4,\cdots,p} \left(1 - 2z^{-1}\cos\omega_i + z^{-2}\right) \tag{5}$$

$$Q(z) = (1 + z^{-1}) \prod_{i=1,3,\cdots,p-1} \left(1 - 2z^{-1}\cos\omega_i + z^{-2}\right) \tag{6}$$

上記の式 (5)，(6) において，p は偶数の場合の LSP の次元数，$\{\omega_i\}$ は LSP となる．また，音声の合成において，安定性を満たす条件式は以下のように定義される．

$$0 < \omega_1 < \omega_2 < \cdots < \omega_p < \pi \tag{7}$$

また，後に説明する SKA による学習で用いる特徴では，標本化周波数を f_s とし，以下の非正規化周波数表現 $\{f_i\}$ を用いる．

$$f_i = \frac{f_s \omega_i}{2\pi} \tag{8}$$

2.2 共役構造-代数符号励振線形予測による特徴抽出

音声から特徴を抽出する方法として，共役構造-代数符号励振線形予測 (CS-ACELP: Conjugate Structure Algebraic Code Excited Linear Prediction)[3] を用いる．特徴抽出の手順は以下の図 1 に示す．まず，時間窓を 30 ミリ秒，時間窓のシフト間隔を 10 ミリ秒とし，各時間窓ごとに特徴を抽出する．具体的には，LP 次数が p 次の場合，各時間窓に対して p 個の LSP パラメータ，音声のパワー (ゲイン)，それらの隣接 5 フレームにおける線形回帰の傾きから求まる，$p+1$ 個の動的特徴量から成る $2(p+1)$ 個の要素を持つベクトルを 1 フレームとする．

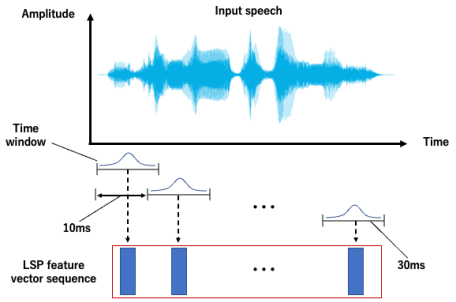

図 1 入力特徴ベクトル系列の抽出

3. 識別関数

3.1 マルチプロトタイプ状態遷移モデル

分類器の単語クラスモデルとして，マルチプロトタイプ-状態遷移モデル (MP-STM: Multi Prototype-State Transition Model) を用いる (図 2)．単語に含まれる各音素は S_h 個の状態をもち，さらに各状態は $I_{h,s}$ 個のプロトタイプをもつ構造になっている．また，H 個の音素の内，h 番目の音素の s 番目の状態の i 番目のプロトタイプを $\boldsymbol{r}_i^{h,s}$ と表し，その要素は入力特徴ベクトルと同様に，LSP パラメータを含む $2(p+1)$ 次元のベクトルとする．また状態遷移モデルに含まれる全てのプロトタイプをまとめて $\boldsymbol{\Lambda} = \{\{\{\boldsymbol{r}_i^{h,s}\}_{i=1}^{I_{h,s}}\}_{s=1}^{S_h}\}_{h=1}^{H}$ とする．

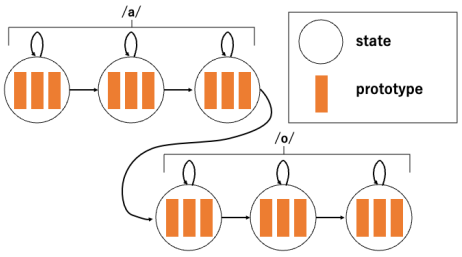

図 2 単語"a o"のマルチプロトタイプ状態遷移モデル
(状態数 $S_h = 3$，プロトタイプ数 $I_{h,s} = 3$)

3.2 動的時間伸縮

音声データから抽出した入力特徴ベクトル系列と単語クラスモデルの類似度を表す識別関数として，動的時間伸縮 (DTW: Dynamic Time Warping) に基づく距離 (図 3) を採用する．

フレーム数が T の入力特徴ベクトル系列 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T]$ とクラス C_j との DTW に基づく最小累積距離は以下のように計算できる．

$$g_j(\boldsymbol{X}; \boldsymbol{\Lambda}) = \frac{1}{T}\sum_{t=1}^{T} \left\| \boldsymbol{x}_t - \boldsymbol{r}_{i(\phi_{j,t},\theta_{j,t},t)}^{\phi_{j,t},\theta_{j,t}} \right\|^2 \tag{9}$$

ここで，t は入力特徴ベクトル系列のフレーム番号，$\phi_{j,t}$ および $\theta_{j,t}$ はフレーム t の入力特徴ベクトル $\boldsymbol{x}_t$ に対する，単

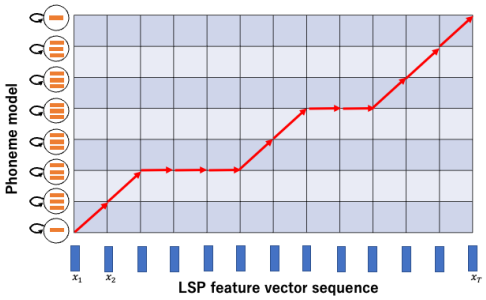

図 3　DTW に基づく識別関数計算のイメージ図

語クラス C_j の状態遷移モデル内の最近傍プロトタイプを含む音素と状態のインデックスを表わす．$i(\phi_{j,t}, \theta_{j,t}, t)$ は，音素 $\phi_{j,t}$ の状態 $\theta_{j,t}$ における，$\boldsymbol{x}_t$ との最近傍プロトタイプのインデックスであり，以下の式で求まる．

$$i(\phi_{j,t}, \theta_{j,t}, t) = \underset{i=1}{\overset{I_{\phi_{j,t},\theta_{j,t}}}{\arg\min}} \left\| \boldsymbol{x}_t - \boldsymbol{r}_i^{\phi_{j,t},\theta_{j,t}} \right\|^2 \qquad (10)$$

この時，入力特徴ベクトル系列 $\boldsymbol{X}$ の各フレームに対する単語クラス C_j における最適な音素と状態 $(\phi_{j,1}, \theta_{j,1}), \cdots, (\phi_{j,T}, \theta_{j,T})$ を最適経路とする．

4. Segmental K-means Algorithm

音声データなどの時系列データに対して K 平均法を適用する手法として，Segmental K-means Algorithm (SKA) が存在する．まず SKA では，学習する単語クラスの状態遷移モデルに含まれる，全てのプロトタイプを初期化する．音素の数が H，状態数が S の入力特徴ベクトル系列に対して，各フレームのベクトルの要素である音声パワーを基に無音区間と発話区間に分割する．次に発話区間のベクトル系列を $H \times S$ 個に等分割し，図 4 のように状態遷移モデルの各プロトタイプにベクトル系列を割り当てる．これを，全ての単語音声に対して行い，各プロトタイプに割り当てられたベクトルの平均をとる．

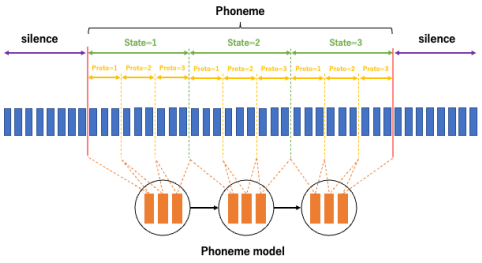

図 4　Segmental K-means Algorithm のプロトタイプ初期化

学習では，初期化した単語クラスモデルを基に DTW を用いて入力特徴ベクトル系列の各フレームを再度，単語クラスモデルの各プロトタイプに割り当て，平均をとる (図 5)．以降は同様の手順をプロトタイプの更新量が閾値以下になる，あるいは，繰り返し上限に達することで学習は終了する．

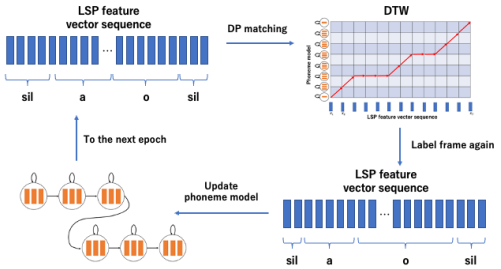

図 5　Segmetal K-means Algorithm の学習手順

5. 成人音声と子供音声の線スペクトル対の分布

5.1 データセット

5.1.1 ETL-WD-II

成人の音声データセットとして，国立情報学研究所が提供する単語音声データベース ETL-WD-II を用いる [7]．このデータセットは男女各 10 名の計 20 名の話者が発話した 1542 語の単語音声データが収録されている．音声データは 16kHz で標本化，16 ビットで量子化されている．本実験では，この内，男女各 5 名の計 10 名，各話者ごとに 492 単語を用いる．

5.1.2 JWC

子供の音声データセットとして，国立情報学研究所が提供する JWC を用いる [8]．このデータセットは男性 13 名，女性 11 名の計 24 名の，10 歳から 11 歳の話者が発話した 653 語の単語音声データが収録されている．音声データは 16kHz で標本化，16 ビットで量子化されている．本実験では，この内男子 6 名，女子 4 名の計 10 名，各話者ごとに 424 単語を用いる．

6. 線スペクトル対と状態遷移モデルのプロトタイプの散布図

6.1 実験概要

本実験では，ETL-WD-II と JWC に対する各音素と各状態における LSP の分布を各次元ごとに可視化し，成人音声と子供音声の LSP の分布の差異を視覚的に確認する．具体的には，ETL-WD-II と JWC に対して，SKA による学習を行い音素モデルを作成する．SKA の学習における実験条件は表 1 に示す条件で行う．次に，学習した音素モデルを基に ETL-WD-II と JWC の全ての音声標本における発話区間のフレームに対し，DTW による最適経路から音素と状態を割り当てる．そして，割り当てられたフレームは各音素の各状態ごとに散布図としてプロットする (図 6 ～ 図 20)．この時，散布図は 2 次元とし，LSP の次元数 (10 次元)

の内から 2 次元選択し縦軸と横軸に設定する．加えて，学習した音素モデルにおけるプロトタイプも同時に散布図へプロットする (図 6 ～ 図 20 中の星印)．

表 1 Segmental K-means Algorithm の実験条件

状態数 (S_h)	3
プロトタイプ数 ($I_{h,s}$)	3
LSP の次元数	10
エポック数	100

6.2 各次元における線スペクトル対の分布

まず，成人音声と子供音声の LSP の各次元における分布を確認する．音素は日本語母音，即ち/a/, /i/, /u/, /e/, /o/を使用し，状態は音素全体の中心部を表わす 2 番目の状態を選択する．また，10 次元の LSP の次元の内，選択された 2 つの次元を散布図の横軸と縦軸とし，LSP の単位は Hz とする．散布図中の各点は，ETL-WD-II と JWC における全ての入力特徴ベクトル系列の内，SKA で学習した音素モデルとの DTW による音素ラベリングで選択された音素の状態 2 である全てのフレームの LSP の値を表す．加えて星型の点は，SKA で学習した音素モデルの状態 2 に含まれる 3 つのプロトタイプ位置を表す．

音素/a/における散布図を図 6 から図 8 に，音素/i/における散布図を図 9 から図 11 に，音素/u/における散布図を図 12 から図 14 に，音素/e/における散布図を図 15 から図 17 に，音素/o/における散布図を図 18 から図 20 に示す．各音素の散布図では，低次元から高次元に推移することで，LSP の分布が，どのように変化するか確認するため，軸を (1 次元目, 2 次元目), (5 次元目, 6 次元目), (9 次元目, 10 次元目) の 3 つに選択した．これらを基に，成人音声と子供音声の母音における LSP の分布を各次元ごとに確認する．

まず，音素/a/, /i/, /u/, /e/, /o/の全てにおいて見られた傾向として，低次元 (1 次元目, 2 次元目) の LSP においては成人音声 (ETL-WD-II) と子供音声 (JWC) の間に分布の差異が生じていることが確認できた．子供音声の方が成人音声に比べて LSP の値が高く，これは子供音声の方がフォルマント周波数が高く，LSP がフォルマント位置に密に生じることからもわかる．また，高次元 (9 次元目, 10 次元目) の LSP においては成人と子供の間にほとんど分布の差は現れなかった．しかし，LSP は音声のフォルマント周波数の両端を挟むように対で生じるため，1 次元目，2 次元目の LSP は第 1 フォルマント周辺に位置し，逆に，9 次元目，10 次元目の LSP は高次のフォルマント周辺に生じることから，低次元における LSP の分布が音声認識においては重要な役割を果たすことがわかる．低次元と高次元の中間にあたる次元 (5 次元目, 6 次元目) では，音素/a/, /u/, /o/では，ほとんど差が見られない，もしくは，微小な分布の差であり，音素/i/, /e/では，大きく分布の差が生じたことから，母音の種類によって異なる結果となった．

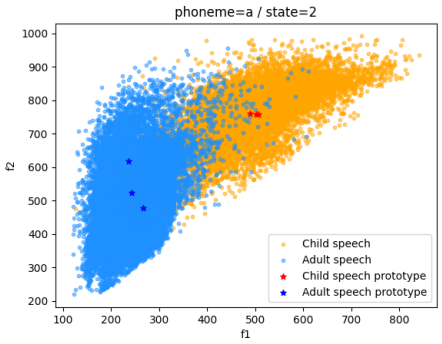

図 6 音素/a/の状態 2 における LSP の分布 (横軸: 1 次元目, 縦軸: 2 次元目)

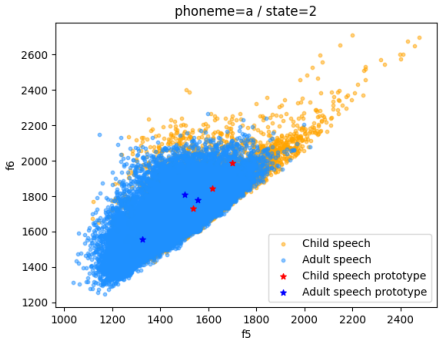

図 7 音素/a/の状態 2 における LSP の分布 (横軸: 5 次元目, 縦軸: 6 次元目)

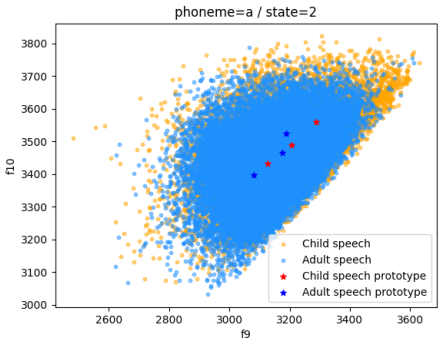

図 8 音素/a/の状態 2 における LSP の分布 (横軸: 9 次元目, 縦軸: 10 次元目)

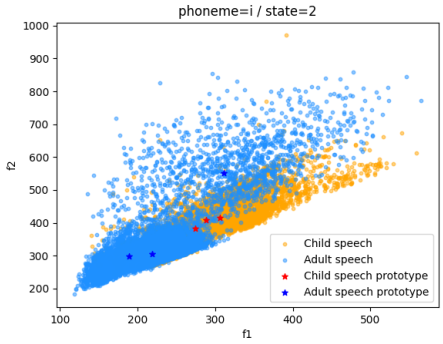

図 9　音素/i/の状態 2 における LSP の分布
(横軸: 1 次元目, 縦軸: 2 次元目)

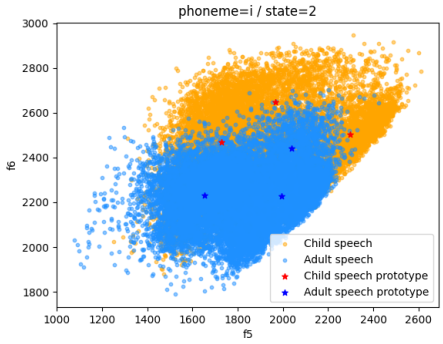

図 10　音素/i/の状態 2 における LSP の分布
(横軸: 5 次元目, 縦軸: 6 次元目)

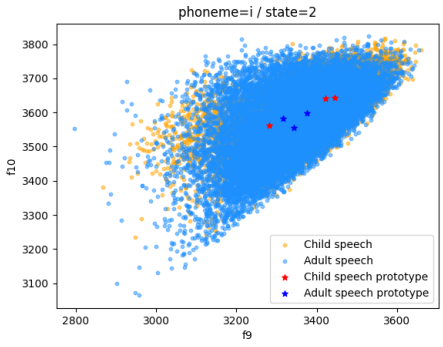

図 11　音素/i/の状態 2 における LSP の分布
(横軸: 9 次元目, 縦軸: 10 次元目)

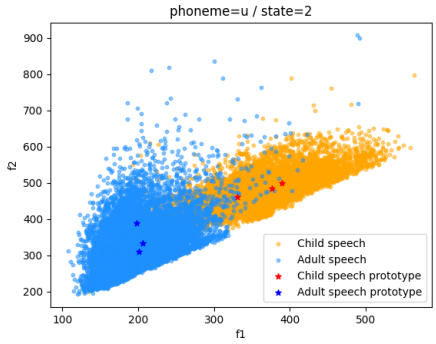

図 12　音素/u/の状態 2 における LSP の分布
(横軸: 1 次元目, 縦軸: 2 次元目)

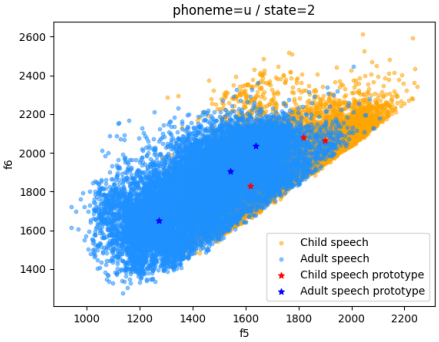

図 13　音素/u/の状態 2 における LSP の分布
(横軸: 5 次元目, 縦軸: 6 次元目)

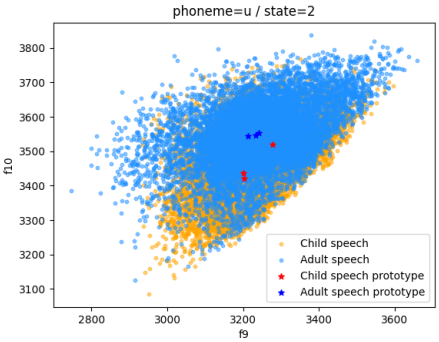

図 14　音素/u/の状態 2 における LSP の分布
(横軸: 9 次元目, 縦軸: 10 次元目)

7.　おわりに

成人音声と子供音声において, 母音においては比較的低い次元の LSP においては, 分布の差を視覚的に捉えることができた. また, 次元を上げると共に, 分布の差が生じるかどうかは, 母音ごとに異なる結果となった.

今後の課題としては, これらの LSP の特性を子供音声の認識にどのように組み込み精度向上を図るかが挙げられる. また, 他の成人, 子供の音声データに対しても同様の性質が現れるかを検証する必要がある.

謝辞　本研究の一部は, 科研費 18H03266 の支援を受けて行った. 著者一同, 感謝申し上げる.

参考文献

[1] 竹内 幸将. 子供の音声を用いた音声合成可能性を正則化条件とする最小分類誤り学習法の実験的評価. 同志社大学理工学部卒業論文, February, 2019.

[2] Naoto Umezaki, Takumi Okubo, Hideyuki Watanabe, Shigeru Katagiri, Miho Osaki. Minimum Classification Error Training with Speech Synthesis-based Regularization for Speech Recognition. International Conference on Signal Processing and Machine Learning (SPML), November, 2019.

[3] ITU-T Rec. G.729. Coding of Speech at 8 kbit/s us-

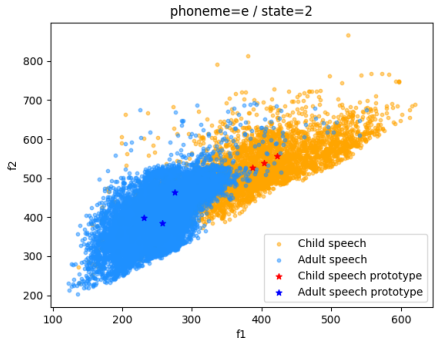

図 15 音素/e/の状態 2 における LSP の分布
(横軸: 1 次元目, 縦軸: 2 次元目)

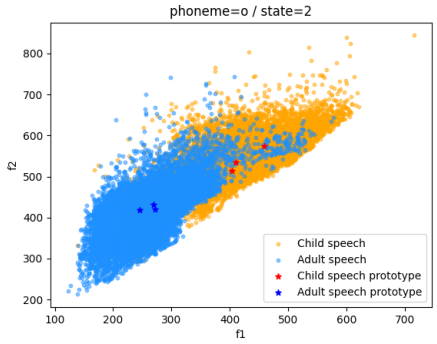

図 18 音素/o/の状態 2 における LSP の分布
(横軸: 1 次元目, 縦軸: 2 次元目)

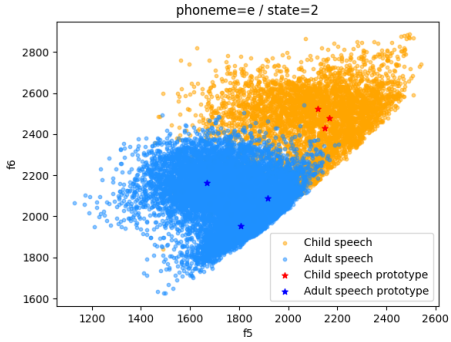

図 16 音素/e/の状態 2 における LSP の分布
(横軸: 5 次元目, 縦軸: 6 次元目)

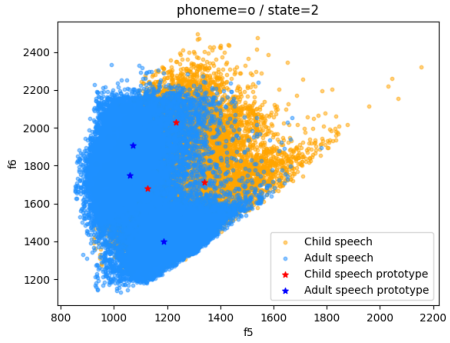

図 19 音素/o/の状態 2 における LSP の分布
(横軸: 5 次元目, 縦軸: 6 次元目)

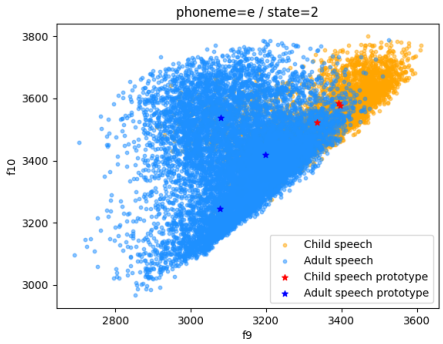

図 17 音素/e/の状態 2 における LSP の分布
(横軸: 9 次元目, 縦軸: 10 次元目)

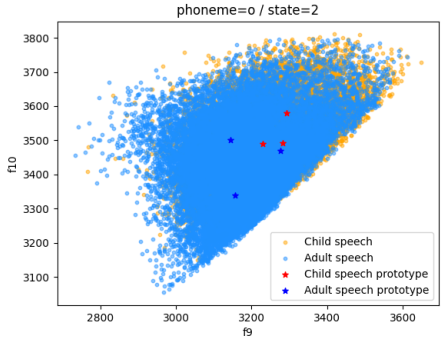

図 20 音素/o/の状態 2 における LSP の分布
(横軸: 9 次元目, 縦軸: 10 次元目)

ing Conjugate Structure Algebraic Code Excited Linear Prediction(CS-ACELP).

[4] Biing-Hwang Juang and Lawrence R Rabiner. The segmental k-means algorithm for estimating parameters of hidden markov models. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 38, pp.1639-1641, No.9 1990.

[5] F. Itakura. Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals. Journal of Acoustical Society of America, Vol. 57, S35, 1975.

[6] F. Itakura, T. Kobayashi, and M. Honda. A Hardware Implementation of a New Narrow to Medium Band Speech Coding. Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1982, pp. 1964-1967, May 1982.

[7] https://doi.org/10.32130/src.ETL-WD

[8] https://doi.org/10.32130/src.JWC