ヒトの聴覚末梢系モデルを用いた音声情報処理

山口大学大学院理工学研究科　システム情報科学研究室

背景と目的

音声処理システム

強力な統計モデルに基づいて音声処理を行っているものがほとんどである

大きな騒音などがある場合には対応できない場合がある

ヒトの聴覚系

環境や距離，媒体が違っていても必要な情報のみに注目して会話ができる

様々な環境に対応できる

ヒトの聴覚系のような柔軟性をもつ音声処理システムの開発が望まれている

背景

ヒトの聴覚系のような柔軟性をもつ音声処理システムの開発が望まれている

本研究の目的

目的

・ヒトの聴覚系の一部である聴覚末梢系の数理モデルを構築する
・聴覚末梢系モデルの音声処理システム（話者識別）への応用を考える

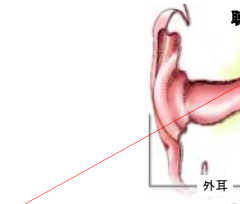
本研究で構築するモデルとその応用

入力が音声信号，出力がパルス時系列信号である聴覚末梢系モデルを構築する

モデルの出力であるパルス時系列信号に基づく特徴量を用いて話者識別を行う

話者識別精度の向上を図るための特徴量変換を行う

聴覚末梢系モデル



聴覚末梢系モデル

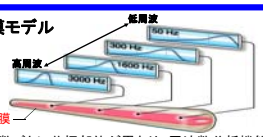
① 基底膜モデル → ②-i 順応特性モデル → ②-ii 位相固定性モデル → ③ 聴神経モデル

（②：有毛細胞モデル）

音声信号 ― 周波数スペクトル ― 発火確率 ― パルス時系列信号

入力：音声信号 → モデル → 出力：パルス時系列信号

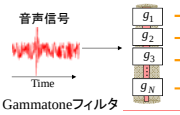
① 基底膜モデル



基底膜･･･周波数ごとに共振部位が異なり，周波数分析機能を有する

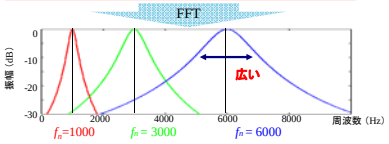
周波数分析･･･フィルタバンクで行うことができる

Gammatone フィルタバンクを使用して基底膜のモデル化を行う

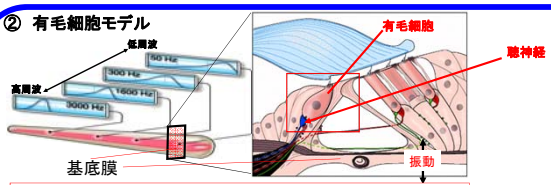


Gammatoneフィルタ

$$g_n(t) = at^3 \exp(-2\pi bt)\cos(2\pi f_n t + \phi) \qquad (n = 1,2,\cdots,N)$$

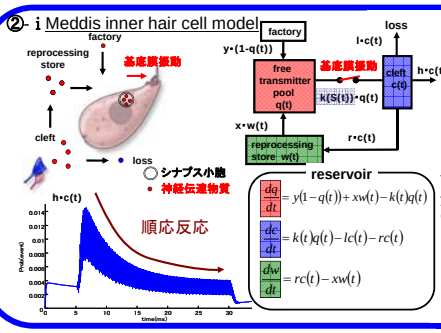


② 有毛細胞モデル



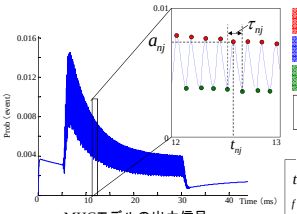
有毛細胞の振動変換機能をMeddis inner hair cell modelで表現する

②- i Meddis inner hair cell model



②- ii 位相固定性モデル

Meddis inner hair cell modelでは表現できない位相固定性の消失を解決するため，低域通過フィルタを用いずに，牧らのモデルを導入する



MHCモデルの出力信号

a_{nj} ：極大点の振幅値（振幅情報）

t_{nj} ：時間（位相情報）

τ_{nj} ：極大値に隣接する極小点間の時間間隔（周波数情報）

$$\{a_{nj}, t_{nj}, \tau_{nj}\} \quad (j = 1,2,\cdots,J_n - 1)$$

揺らぎを加える

位相固定性の消失をモデル化

$$t'_{nj} = t_{nj} + N\{0, f(\tau_{nj}) \cdot s(a_{nj})\}$$
$$f(\tau_{nj}) = \tau_{nj}(\omega_1 \tau_{nj}^{-\omega_2} + \omega_3)$$
$$s(a_{nj}) = 1 + r_1 e^{-r_2 a_{nj}}$$

有毛細胞モデルの出力

$$\{a_{nj}, t'_{nj}, \tau_{nj}\} \quad (j = 1,2,\cdots,J_n - 1)$$

③ 聴神経モデル

聴神経の膜電位は神経伝達物質により上昇しその後，減衰する

単純な関数 $t \cdot \exp(-t)$ を用いてモデル化する

時刻 t'_{nj} における膜電位を振幅値 a_{nj} に比例させる

$$a_{nj}(t - t'_{nj}) \cdot \exp\{-(t - t'_{nj})\}$$

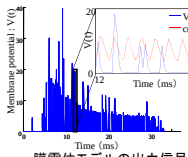
膜電位のモデル化

$$V(t) = \sum_{\{j | t'_{nj} + h(\tau_{nj}) < t\}} u a_{nj} \left[t - t'_{nj} - h(\tau_{nj})\right] \cdot \exp\left(-\left[t - t'_{nj} - h(\tau_{nj})\right] T(\tau_{nj})\right)$$
$$h(\tau_{nj}) = h_1 + h_2 \tau_{nj} \qquad T(\tau_{nj}) = T_1 \cdot \exp(-T_2 \tau_{nj})$$

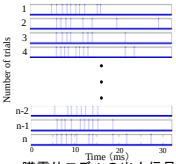
膜電位 $V(t)$ がしきい値以上のときに発火，ただし，不応期の間は発火しない

パルス生成のモデル化

$$S(t) = \begin{cases} 1 & V(t) \geq U(\alpha, \beta) \\ & \text{and } S(t') = 0 \quad \text{for } t' \in [t - t_r, t] \\ 0 & \text{otherwise} \end{cases} \qquad t_r \sim N(\mu_r, \sigma_r^2)$$



膜電位モデルの出力信号



膜電位モデルの出力信号

検証実験

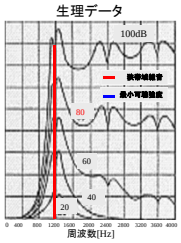
構築したモデルからヒトの聴覚系が有する聴神経の特性を検証する

主な2つを紹介する

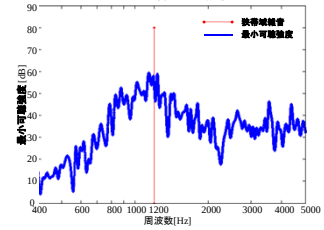
マスキング

ある音の最小可聴域が他の音の存在によって上昇する現象

生理データ



モデルから得られた結果

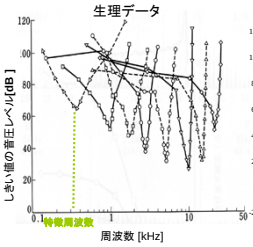


構築したモデルで聴覚系で見られるマスキングを再現できていることがわかる

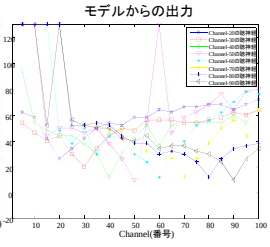
周波数選択性

ある聴神経はある特定周波数の音に対して発火しやすい現象

生理データ



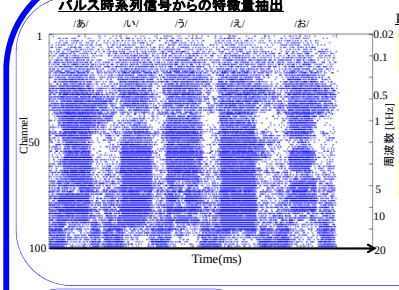
モデルからの出力



構築したモデルで聴覚系で見られる周波数選択性を再現できていることがわかる

話者識別実験

パルス時系列信号からの特徴量抽出

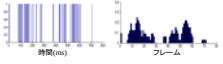


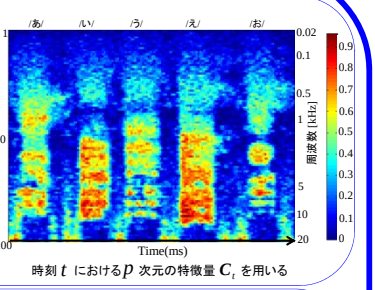
Post-Stimulus Time Histogram(PSTH)

パルス列のラスター表示において，時間軸上を一定幅の小区間で分割

各区間に入っているパルス数を試行全体にわたって平均化

この平均値を区間の代表値としてヒストグラムを作成





時刻 t における p 次元の特徴量 C_t を用いる

特徴量変換

話者識別率を向上させるため，特徴量変換を行う

1. **標準化変換** $C_t \rightarrow \tilde{C}_t$
特徴量を，平均値および共分散行列を用いて標準化する。但し，共分散行列は対角成分のみをもつものとして扱う。

2. **正規化変換** $C_t \rightarrow \hat{C}_t$
特徴量をそのノルムで正規化する

3. **標準・正規化変換** $C_t \rightarrow \hat{\tilde{C}}_t$
標準化後に正規化変換を行う

話者識別実験

男性9名，女性3名の母音発話各5回をモデルに入力し，特徴量 C_t を求める

特徴量変換により，それぞれの特徴量 $\tilde{C}_t$ ，$\hat{C}_t$ ，$\hat{\tilde{C}}_t$ を求める

最近隣法を用いて，5 hold leave 2 out cross-validation により話者識別率の評価を行う

比較のためメルLPCスペクトルを特徴量とする話者識別実験も行った

話者識別結果

話者識別率の平均値[%]

	無変換	標準化	正規化	標準・正規化
LPC	80.0	78.1	84.8	80.6
	(3.42)	(3.96)	(2.28)	(2.52)
Proposed method	78.1	86.1	79.8	86.6
	(3.47)	(2.12)	(2.28)	(3.79)

( )は標準偏差を表す

今後の課題

多次元パルス信号から得られる様々な特徴量を用いて，さらなる話者識別率の向上を目指す