聴覚情景物体生成過程の知見を用いた重畳音声分離向けラダーネットワークの動作解析

〇関口　浩　　成末　義哲　　森川　博之
東京大学大学院　工学系研究科

1. はじめに

複数の話者の重畳音声から目的の音声を抽出する音源分離と再構成は，議事自動生成や聴覚支援等への応用が期待されている．本研究では聴覚脳神経学の知見を基に，音源分離機能の数理モデルを導き，複数種類のラダーネットワークの組み合わせにより実装することで，音源分離と再構成を行う．

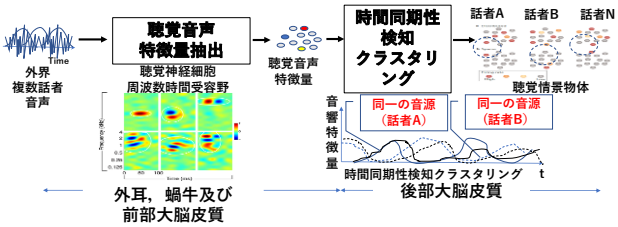
2. 聴覚脳神経学の知見

人類は同時に発声する外部音源を聴覚脳神経系で聞き分ける機能がある．

■聴覚音声特徴量抽出：聴覚神経細胞の周波数時間受容野で音声を分析し聴覚音声特徴量時間系列を出力する機能．

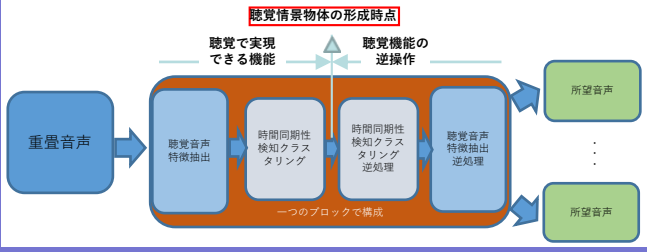
■時間同期性検出クラスタリング：音声特徴量時間系列の発生開始時間点および終了間点が類似のものを同一の音源とし，異るものを他の音源として判断する機能．

- 開始時間と終了時間はそれぞれ5Hz 以下の緩やかな動きである話者の唇の開口点と閉口点に相当．
- 同一の音源からの特徴量集団を聴覚情景物体と呼ぶ．



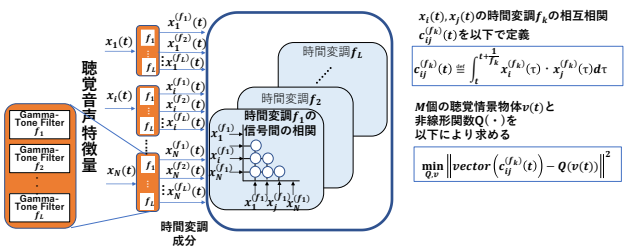
3. 全体構成

聴覚２機能とその逆処理を直結する．



4. 数理モデル

■時間同期検出クラスタリング動作の実現
→時間コヒーレントで数理モデル化
唇の動きを検知



■聴覚音声特徴量と時間同期検知クラスタリングとの親和性の実現
→聴覚音声特徴量の相互の生成確率が独立で，かつ時間方向の自己相関を保持
→非線形スパースエンコーダデコーダで数理モデル化

$$\min_{J,s} \sum_t \left(\underbrace{\|X(f,t)\| - J(s(t))\|^2}_{\text{エラー項}} + \underbrace{\lambda \|s(t)\|_1}_{\text{sparse条件}} \right)$$

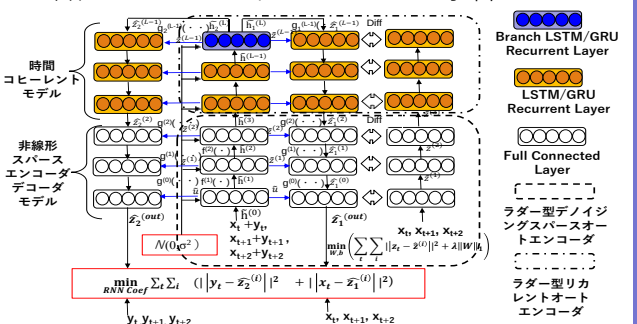
$\|X(f,t)\|$：１話者の短時間音声スペクトラム
$s(t)$：隠れ層変数ベクトル
$J(*)$：非線形関数
$\|\cdot\|_1$ ℓ１ノルムで$s(t)$のsparse正則化条件
λ：ハイパーパラメータ

5. ラダーネットワーク実装

■再構成パスの確保のためにオートエンコーダの1種であるラダーネットワークを採用

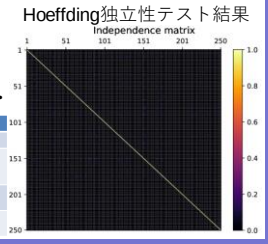
■ LSTM と Full Connected のラダーネットワークを重ねて分離再構成を実現

■音声データからネットワークを学習



6. 初期動作解析

■聴覚音声特徴量は互いに独立であることが示された．



項目	パラメータ
音声コンテンツ	男女各１名の音声(各約20分，合計約40分) RWCPニュース原稿
学習データ（男女各17分）教師無し学習	短時間フーリエ変換振幅: 8KHzサンプリング, FFT:512point, 切り出し窓: 256point(32msec)，窓シフト: 128point(16msec) 257point/frame x 4 frame = 1028point/データ，34,638データ/epoch
ネットワークモデル	(全結合，Batch Normalization，Relu)/階層，階層：6層 次元数: (一層)2000，1000，1000，500，500，(6層)250
テストデータ（男女各3分）	フォーマットは学習データと同じ，5,965データ/epoch

連絡先：　関口　浩　　sekigh@mlab.t.u-tokyo.ac.jp