多人数多マイクでの発話区間検出 ～ピンマイクでの事例～*

○澤田 宏，荒木 章子，大塚 和弘，藤本 雅清，石塚 健太郎 (NTT 研究所)

1 はじめに

発話区間検出は重要な技術であり，適応フィルタの学習制御に用いられたり，音声認識の湧き出し誤りの低減に寄与したり，収録された会議音声のイベント構造化などにも役立つ．これまでに，単一のマイクロホンを用いて，話者一人を仮定した発話区間検出法が多く提案されてきた（例えば [1]）．

本稿では，図 1 のような，話者が複数人居て各話者の胸元にピンマイクが装着されているような状況を考え，それぞれの話者の発話区間を検出することを課題とする．ピンマイクを装着しているという音響的に良好な状況でも，多人数が集まって会話をすると，図 2 に示すように，他人の声が自分のピンマイクに入り込むため，従来の単一話者を仮定した発話区間検出法はあまり精度が出ない（図 4 に検出結果を示す）．そこで本稿では，複数マイクでの時間周波数成分を話者毎の発話に分類することで，高精度に発話区間を検出する方法を提案する．

2 発話区間検出技術

まず，単一のマイクロホンを用いて発話を検出する従来技術の一つ [1] を説明する．マイクロホンでの観測信号に短時間フーリエ変換を施し，時間周波数表現 $x(f,t)$ を得る．また，周波数 f 毎にノイズパワーの推定値 $\lambda(f)$ を，例えば最初の何秒間かは発話がないと仮定するなどして，何らかの手段で得る．そして，各時間周波数毎に事後 S/N 比

$$\gamma(f,t) = \frac{|x(f,t)|^2}{\lambda(f)} \qquad (1)$$

を計算し，以下の式に従って，ある種の非線形変換を施した後，考慮するすべての周波数で平均化する．

$$G(t) = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} [\gamma(f,t) - \log \gamma(f,t) - 1] \qquad (2)$$

$\mathcal{F}$ は考慮する周波数の集合であり，$|\mathcal{F}|$ は集合 $\mathcal{F}$ の要素の数である．このように計算した $G(t)$ が閾値 η より大きければ，時間 t でのフレームは発話区間，小さければ非発話区間であると判定する．なお，ここでの非線形変換は，観測信号 $x(f,t)$ をノイズと発話に分類して，それぞれを分散の異なるガウス分布でモデル化した際の尤度比から導出されるものである [1]．

図 1 の条件で収録した 4 つのピンマイクのそれぞれの観測信号（図 2）に対して，単一マイクを仮定した本従来手法を適用したところ，図 4 の上段に示すような結果が得られた．他人の発話まで過剰検出してしまっていることが見てとれる．

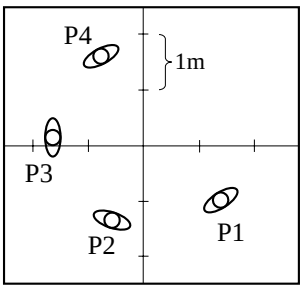

図 1: 実験条件．4 人の話者，胸元にピンマイク．

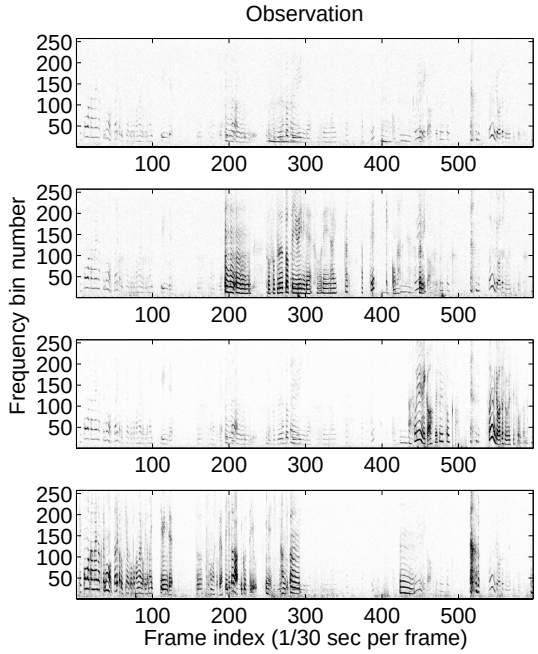

図 2: 各話者のピンマイクでの観測信号．時間周波数表現に変換されている．

3 提案手法

提案手法では，まず，複数マイク（M 本）での観測信号の時間周波数表現をベクトル $\mathbf{x}(f,t) = [x_1(f,t), \ldots, x_M(f,t)]^T$ にして考える．そして，従来手法と同様に時間周波数毎の観測信号を発話／ノイズに分類した後，さらに，発話部分を話者毎に分類する．分類結果は，0 から N（話者数）までの値をとり得るクラスタ情報 $C(f,t)$ で表現する．ある時間周波数スロット (f,t) において，$C(f,t) = 0$ であれば，そのスロットの観測信号ベクトル $\mathbf{x}(f,t)$ はノイズと分類され，$C(f,t)$ が 1 から N までの値 i を取れば，i 番目の話者による発話と判定されたことを意味する．

本稿で考慮するようなピンマイクを用いた状況では，発話がそれぞれのマイクロホンにどの程度の音量比で観測されたかという情報に基づいて，話者毎

*Voice activity detection for multiple speakers with multiple lavalier microphones, by SAWADA Hiroshi, ARAKI Shoko, OTSUKA Kazuhiro, FUJIMOTO Masakiyo, ISHIZUKA Kentaro (NTT Corporation)

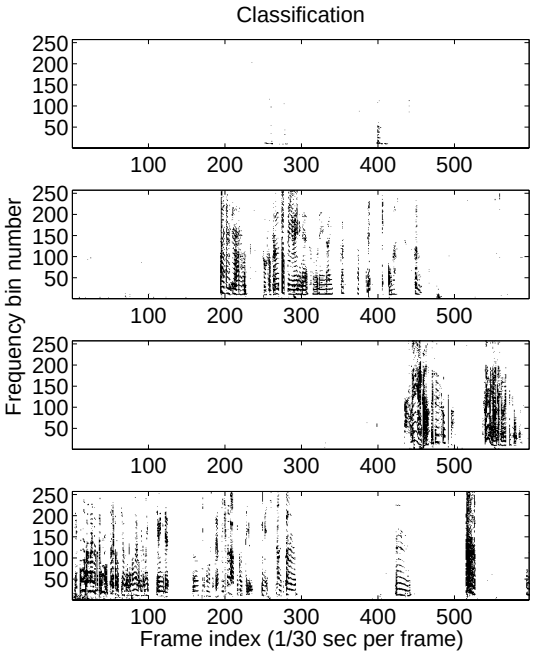

図 3: 提案手法における各話者への分類結果

への分類を行う．そのため，絶対値のみを考慮しノルムを 1 に正規化した新たなベクトル

$$\bar{\mathbf{x}}(f,t) = \begin{bmatrix} \bar{x}_1(f,t) \\ \vdots \\ \bar{x}_M(f,t) \end{bmatrix}, \quad \bar{x}_i(f,t) = \frac{|x_i(f,t)|}{||\mathbf{x}(f,t)||} \tag{3}$$

を考える．このベクトル $\bar{\mathbf{x}}$ は，話者毎にクラスタを形成することが期待できる．なぜなら，各話者 i の近くにピンマイク i が配置されている状況を考慮しているからである．例えば話者 1 のクラスタには，$\bar{x}_1(f,t)$ が相対的に大きな値を持つようなベクトル値が集まる．

さて，話者 i の発話に対応するクラスタを，平均ベクトル $\mathbf{m}_i$，共分散行列 $\sigma_i^2\mathbf{I}$ の多次元ガウス分布

$$p_i(\bar{\mathbf{x}}) = \frac{1}{(\sqrt{2\pi}\sigma_i)^M} \exp\left(-\frac{||\bar{\mathbf{x}} - \mathbf{m}_i||^2}{2\sigma_i^2}\right) \tag{4}$$

でモデル化すると，各時間周波数スロットでのベクトル $\bar{\mathbf{x}}(f,t)$ に対して，

$$C(f,t) = \mathrm{argmax}_i\, p_i(\bar{\mathbf{x}}(f,t)) \tag{5}$$

により，分類結果 $C(f,t)$ を得ることができる．各話者に対応する平均ベクトル $\mathbf{m}_i$ は，ピンマイクの状況を考慮すると適切に設定できる．例えば，話者 1 については，$\mathbf{m}_1 = [1,0,\ldots,0]^T$ と設定する．その後，平均ベクトル $\mathbf{m}_i$，分散 σ_i^2 共に，適宜更新していくことが望ましい．

図 2 に示す観測信号を上記の方法で各話者に分類した結果を図 3 に示す．各時間周波数スロット (f,t) では高々一人の発話者を仮定しているが，時間 t のフレームは複数の周波数スロットを持つため，同一フレーム内で複数話者の発話を許すモデルとなっている．

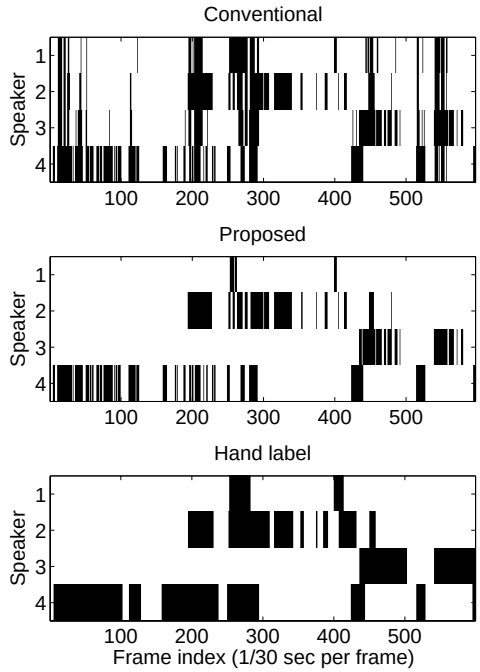

図 4: 発話区間検出結果．上から，従来手法，提案手法，人手によるラベル．

最後に，分類結果に基づいて，各話者 i に対応する分離信号 y_i を

$$y_i(f,t) = \begin{cases} x_i(f,t) & \text{if } C(f,t) = i, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

として算出し，式 (1) における $x(f,t)$ を $y_i(f,t)$ に置き換えた後，話者 i 毎の発話区間の判断を従来手法と同様に行う．

図 4 の中段に，提案手法による発話区間検出結果を示す．人手による正解ラベルとほぼ同等の結果が得られた．なお，式 (2) に対する閾値 η は，従来手法と提案手法で同じ値を用いた．提案手法では，式 (5) による分類と式 (6) による時間周波数マスキングにより，ピンマイクへの他人の声の混入を抑圧し，過剰検出を抑えていることが分かる．

4 おわりに

多人数でのピンマイク収録に適した発話区間検出法を提案した．時間周波数毎に観測信号ベクトルをノイズと各話者に分類することがポイントである．その点において本手法は，音声のスパース性に基づく音源分離手法 [2] と関連が深い．

参考文献

[1] J. Sohn *et al.* “A statistical model-based voice activity detection,” *IEEE Signal Processing Letters*, 6(1), pp. 1–3, 1999.

[2] 荒木他，“観測信号ベクトル正規化とクラスタリングによる音源分離手法とその評価,” 音講論（秋）2-2-3, pp. 591–592, 2005.