発話の連続性に基づいた音声信号の分類による会議音声の可視化*

加藤通朗，☆杉本侑哉，宮部滋樹，牧野昭二，山田武志，北脇信彦 (筑波大)

1 はじめに

近年，会議の様子を収録した音声や動画をコンテンツとしてアーカイブ化するにあたり，話者の発話位置や発話時間を検出する技術（話者インデックス化），参加者間のやり取りを検出する技術や，会議の情景を効率的に伝送・再生する技術が活発に研究されている．話者インデックス化では例えば，Araki ら [1] による少数のマイクロホンを用いて各時刻の発話者を同定し音声強調を行う手法，Pardo ら [2] によるマイクロホン間の距離のみを事前情報として用いた話者方向の検出，Anguera ら [3] による会議室内にある複数マイクを用いた，ビームフォーミングによる話者の検出などが挙げられる．また我々はこれまで，上記のような先行研究と同様に，マイクロホンアレーによる会議コンテンツの収録システムと，話者インデックス付きの再生システムを開発している [4].

このシステムの発展系として我々は，会議録再生支援のための会議全体の可視化フレームワークであるカンバセーション・フローを提唱する．これは，音声の到来方向情報による話者の時間的・空間的な情報に加え，連続音声（continuous），断続音声（intermissive），突発音声（impulsive）といった発話状態の情報を，連続発話，相槌や返答に対応するように分類・付与し，会議全体を圧縮表現により可視化したものである．視認性を高めるために各発話の音声クラスに応じた色付けを行っており，これによって閲覧ユーザーは会議の全体を俯瞰でき，会議中で重点的に再生したい部分を容易に見つけることが可能となる．また,「うんうん」「はい」といった相槌や返答は，対話の成り立ちや話題の展開を察する上で重要な情報である一方，長い会議の中ではごく短いものであり，通常の発話インデックスをそのままで圧縮して表現すると，見逃されてしまうことがある．このことからも，これらを「断続音声」「突発音声」として予め検出し，色付けによって強調しておくことは有用である．

本論文では，カンバーセーション・フローを実現するための，連続音声・断続音声・突発音声を分類して検出する手法を提案する．マイクロホンアレーで観測された信号を MUSIC やビームフォーマなどで分析し，これらの発話状態の特徴が方向推定スコアの時間変化に強く表れることに着目する．各方向の方向推定スコアは，発話が連続的なほど，なだらかな変化を取り，また発話が突発的であるほど急速に時間変化するため，この特徴を方向ごとの短時間フーリエ分析を用いて求める．これにより得られる短時間振幅スペクトルから，フィルタバンク分析により手動でパラメータを調整して発話状態を分類する手法と，サポートベクターマシン（Support Vector Machine; SVM）などの分類器で自動分類させる手法を提案する．

2 収録音声からの MUSIC スコアの算出

マイクロホンの数を M，時刻を t としたとき，各マイクロホンで観測される信号は $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_M(t)]^T$ と表記される．ただし T は転置である．この観測信号の短時間フーリエ変換は

$$\mathbf{x}(\omega,\gamma) = [x_1(\omega,\gamma), x_2(\omega,\gamma), \ldots, x_M(\omega,\gamma))]^T \quad (1)$$

と表される．ここで ω は角周波数，γ は短時間フーリエ変換のフレーム番号，$x_m(\omega,\gamma)$ は m 番目のマイクロホンの観測信号の複素スペクトルである．

このようにフレームごとに得られる観測信号の複素スペクトルについて，ある一定のフレーム単位でまとめたものをブロックと定義し，ブロック単位での処理を考える．b 番目のブロック内のフレーム数を K，ブロック内のフレーム番号を $A_b = \{\gamma_i, \ldots, \gamma_{i+K-1}\}$ としたとき，このブロックの中での空間相関行列は

$$\mathbf{R}_{xx}(\omega,b) = \sum_{\gamma\in A_b} \mathbf{x}(\omega,\gamma)\cdot\mathbf{x}(\omega,\gamma)^H \quad (2)$$

と計算される．ただし H は複素共役転置である．この空間相関行列の固有値展開を考えると，次のように対角化される．

$$\begin{aligned}&\mathbf{V}(\omega,b)^H\mathbf{R}_{xx}(\omega,b)\mathbf{V}(\omega,b)\\ &= \mathrm{diag}[\sigma_1(\omega,b),\ldots,\sigma_N(\omega,b),0,\ldots,0] + \sigma^2 I\end{aligned} \quad (3)$$

ここで $\mathrm{diag}[\sigma_1(\omega,b),\ldots,\sigma_N(\omega,b),0,\ldots,0]$ は空間相関行列の固有値を大きい順に並べた対角行列であり，σ^2 はノイズである．また $\mathbf{V}(\omega,b) = [\mathbf{v}_1(\omega,b),\ldots,\mathbf{v}_N(\omega,b),\ldots,\mathbf{v}_M(\omega,b)]$ の各列成分は，対角行列中の各固有値番号に対応する固有ベクトルである．上式のように空間相関行列は，空間内に存在する音源数分のパワーと，空間に存在する雑音成分のパワーとに分けることができる．空間内に N 個の音源があると仮定した場合，その音源は第 N 番目までの固有ベクトルが張る部分空間内（signal subspace）に存在し，ノイズ成分は各マイクロホン間の雑音が無相関の場合，N 番目までの signal subspace，および $N+1$ 番目以降の固有ベクトルが張る部分空間（noise subspace）内に均等に出現する．この性質を利用し，マイクロホンアレーの任意の方向 θ における音声の到来を示すスコアは以下のようになる．

$$p(\omega,b,\theta) = \frac{1}{\sum_{i=N+1}^{M}|\mathbf{v}_i(\omega,b)^H\cdot\mathbf{a}(\omega,\theta)|^2} \quad (4)$$

ここで $p(\omega,b,\theta)$ をブロック b，角周波数 ω，アレーからの角度 θ における MUSIC スコアと定義し，各方向で算出する．$\mathbf{a}(\omega,\theta)$ は θ における方向ベクトルである．

---

*Visualizing Conversation Flow of Meeting Recording with Speech Classification Based on Continuousness of Utterance, by Michiaki KATOH, Yuya SUGIMOTO, Shigeki MIYABE, Shoji MAKINO, Takeshi YAMADA, Nobuhiko KITAWAKI (University of Tsukuba).

この MUSIC スコアを，興味のある周波数 $\omega_j \sim \omega_k$ で考え，総和を求めた

$$\bar{p}(b,\theta) = \sum_{\omega=\omega_j}^{\omega_k} p(\omega, b, \theta) \tag{5}$$

を広帯域 MUSIC スコアとして用いる．

さらに本論文では，この広帯域 MUSIC スコアをブロックごとに算出し，時系列に並べたブロック系列 MUSIC スコアを，

$$\mathbf{p}(\theta) = [p(b_1,\theta), p(b_2,\theta), \ldots, p(b_k,\theta), \ldots] \tag{6}$$

と定義し，方向ごとのベクトルとして考える．

3 提案手法

3.1 MUSIC スコアのフーリエ分析

断続性あるいは突発性の音声を，連続発話，断続発話，突発発話と分類して検出する手法として，2 章で述べたブロック系列 MUSIC スコアの各方向に対し，フーリエ分析を実施することを考える．

(6) 式より，ブロック k から $k+L-1$ までの連続する L ブロック $B = \{b_k, b_{k+1}, \ldots, b_{k+L-1}\}$ を切り出したブロック系列 MUSIC スコアを $\mathbf{p}(b,\theta)$ としたとき，その短時間フーリエ変換は

$$q(l,k,\theta) = |\sum_{b\in B} w(b-b_k) p(b-b_k,\theta) e^{-j\frac{2\pi(b-b_k)l}{L}}| \tag{7}$$

$$\text{where } w(b) = \begin{cases} \frac{1}{2}\sin(2\pi\frac{b-b_k}{L}) - \frac{1}{2} & (0 \le b \le L) \\ 0 & \text{otherwise} \end{cases}$$

となる．ここで $q(l,k,\theta)$ は，ブロック系列 B の先頭ブロック k から $k+L-1$ を用い，方向 θ での短時間フーリエ変換の結果を示す．ただし l をブロック周波数とする．また $w(b)$ はハミング窓である．この短時間フーリエ変換で得られる振幅スペクトルを，連続性特徴スペクトル（Continuous Feature Spectrum; CFS）と呼ぶこととする．

連続音声，断続音声および突発音声のブロック系列 MUSIC スコアと，得られる連続性特徴スペクトルとの間には，次のような関係があると考えられる．

- 連続音声：連続性特徴スペクトルの直流成分に強い特徴が表れる．
- 断続音声：連続性特徴スペクトルの直流成分および低域に特徴が表れる．
- 突発音声：連続性特徴スペクトルの低域から広域にかけ，なだらかな傾斜が表れる．

図 1 は，ブロック系列 MUSIC スコアと，その短時間フーリエ変換により得られる連続性特徴スペクトルの例を，3 種類の音声それぞれで示したものである．連続性特徴スペクトルのパターンを，3 種類の音声のいずれかに分類することを考えると，連続性特徴スペクトルは，直流成分を含む $(L/2+1)$ 個出現するため，$(L/2+1)$ 次の空間での識別問題に帰着する．

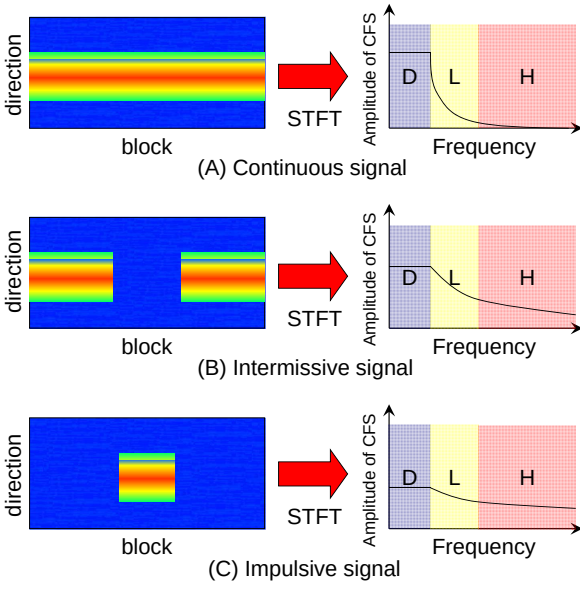

Fig. 1 ブロック系列 MUSIC スコアとフーリエ分析の結果の例

このような問題における SVM による分類では，まず人手によって連続音声，断続音声および突発音声の 3 クラスの正解ラベルが付与されたブロック系列 MUSIC スコアを用いて学習モデルを作成する．さらに，学習モデルと連続性特徴スペクトルのパターンから 3 つのクラスのいずれかに自動分類する．

この手法は，ラベリングを行う人手（ラベラー）の主観的なラベル付与基準を，SVM によって学習することに他ならない．

3.2 フィルタバンク分析による手動分類

フィルタバンク分析による手法では，図 1 に示す通り，$(L/2+1)$ 次の連続性特徴スペクトルを，直流成分（D），低周波領域（L），高周波領域（H）の 3 つの帯域に分割する．各帯域での特徴量を $b_{\mathrm{DC}}, b_{\mathrm{low}}, b_{\mathrm{high}}$ とし，

$$b_{\mathrm{DC}}(k,\theta) = |q(0,k,\theta)| \tag{8}$$

$$b_{\mathrm{low}}(k,\theta) = \sum_{l=1}^{l_{\mathrm{low}}} |q(l,k,\theta)| \tag{9}$$

$$b_{\mathrm{high}}(k,\theta) = \sum_{l=l_{\mathrm{low}}+1}^{l_{\mathrm{high}}} |q(l,k,\theta)| \tag{10}$$

で定義し，それぞれを帯域特徴量と呼ぶこととする．

さて帯域特徴量は，あるブロック系列での連続性特徴スペクトルの各帯域での特徴を表しているだけであり，信号の種類を一意に特定しているわけではない．このため帯域特徴量に対し，3 クラスの音声の類似度ベクトルを別途に定義して，その内積により類似度を算出することを考える．すなわち，連続音声を表す類似度ベクトルを $\mathbf{d}_{\mathrm{cont}}$，断続音声を表す類似度ベクトルを $\mathbf{d}_{\mathrm{int}}$，突発音声を表す類似度ベクトルを $\mathbf{d}_{\mathrm{imp}}$

とし，

$$\mathbf{d}_{\mathrm{cont}} = [d_{\mathrm{cont,DC}}, d_{\mathrm{cont,low}}, d_{\mathrm{cont,high}}] \quad (11)$$

$$\mathbf{d}_{\mathrm{int}} = [d_{\mathrm{int,DC}}, d_{\mathrm{int,low}}, d_{\mathrm{int,high}}] \quad (12)$$

$$\mathbf{d}_{\mathrm{imp}} = [d_{\mathrm{imp,DC}}, d_{\mathrm{imp,low}}, d_{\mathrm{imp,high}}] \quad (13)$$

とする．最終的にいずれかのクラスに分類するための類似度は，これら類似度ベクトルと各帯域特徴量との内積により，スカラー値として

$$E_{\mathrm{cont}} = \mathbf{d}_{\mathrm{cont}} \cdot [b_{\mathrm{DC}}(k,\theta)\ b_{\mathrm{low}}(k,\theta)\ b_{\mathrm{high}}(k,\theta)]^T \quad (14)$$

$$E_{\mathrm{int}} = \mathbf{d}_{\mathrm{int}} \cdot [b_{\mathrm{DC}}(k,\theta)\ b_{\mathrm{low}}(k,\theta)\ b_{\mathrm{high}}(k,\theta)]^T \quad (15)$$

$$E_{\mathrm{imp}} = \mathbf{d}_{\mathrm{imp}} \cdot [b_{\mathrm{DC}}(k,\theta)\ b_{\mathrm{low}}(k,\theta)\ b_{\mathrm{high}}(k,\theta)]^T \quad (16)$$

と定義される．ここで類似度のいずれも，ある閾値より小さい値を取るなら無音とし，それ以外の場合は類似度が最大のクラスに音声を分類するように，下記のルールを定める．

if
$\mathrm{argmax}(E_{\mathrm{cont}}(k,\theta), E_{\mathrm{int}}(k,\theta), E_{\mathrm{imp}}(k,\theta))$
$< E_{\mathrm{threshold}}$
then class = **none**,
else if
$\mathrm{argmax}(E_{\mathrm{cont}}(k,\theta), E_{\mathrm{int}}(k,\theta), E_{\mathrm{imp}}(k,\theta))$
$= E_{\mathrm{cont}}(k,\theta))$
then class = **continuous**,
else if
$\mathrm{argmax}(E_{\mathrm{cont}}(k,\theta), E_{\mathrm{int}}(k,\theta), E_{\mathrm{imp}}(k,\theta))$
$= E_{\mathrm{int}}(k,\theta))$
then class = **intermissive**,
else if
$\mathrm{argmax}(E_{\mathrm{cont}}(k,\theta), E_{\mathrm{int}}(k,\theta), E_{\mathrm{imp}}(k,\theta))$
$= E_{\mathrm{imp}}(k,\theta))$
then class = **impulsive**.

このように，帯域特徴量を信号の性質を定量的に記述したものとし，これに手動で類似度ベクトルを調整することで，検出される断続音声，突発音声が変化する．すなわち，基準として与えられた類似度ベクトルの値を変化させることで，会議コンテンツの閲覧ユーザーの好みに応じて，じっくり聴取する，まばらに聴取するなど，ユーザーの事情に応じたカンバセーション・フローを提供することが可能となる．

4 実験

4.1 会議音声の収録条件と実験条件

会議音声の収録は，7.5m 四方の会議室にて行った．会議室の中央の 1.8m × 1.2m の大きさのテーブルの中心に，直径 20cm の円形マイクロホンアレーを設置した．マイクロホンアレーの円周上には等間隔に 8 つのマイクロホンが搭載されている，サンプリング周波数は 16kHz である．

マイクロホンアレーが設置されているテーブルの周りを 3 人の参加者が，いずれの会議でも 0° 方向，180° 方向および 270° 方向に位置付くように椅子に座っている．話者の移動はないものとする．マイクロホンアレーの中心と，各話者との距離は概ね 1m である．

このようにして，セッションあたり約 15〜20 分程度の会議音声を 4 セッション収録した．このセッションへのラベル付与として，$L = 10$ ブロックの音声について中央 $L/2 = 5$ ブロックでの聴取結果を，ラベラーの主観により判別してラベリングし，その後 $L/2$ ブロックずつシフトしながら全区間を通して行った．

連続性特徴スペクトルは，2 章でのブロック系列 MUSIC スコアを求め，次いで 3.1 節で述べたフーリエ分析を行うことで算出する．Wide-band MUSIC の各ブロックは 0.5 秒とし，ブロック内での短時間フーリエ変換を行うフレーム長は 512，フレームシフトは 128 とした．また空間相関行列を求める際の周波数帯域を 1〜4kHz とした．

このようにして求めたブロック系列 MUSIC スコアを，ラベリングと同様に $L = 10$ ブロックずつ用いてフーリエ分析を行った．得られる連続性特徴スペクトルのビン数は $L/2 + 1 = 6$ となる．ただし直流成分を含んでいる．またフレームシフトも同様に $L/2$ ブロックである．

4.2 分類結果

フィルタバンク分析による手法では，3.2 節で述べた 3 種類の帯域特徴量 $b_{\mathrm{DC}}, b_{\mathrm{low}}, b_{\mathrm{high}}$ を用いて，4 セッションの会議のうち 1 セッションの結果から，類似度ベクトルの手動調整による会議音声の信号の分類を行った．手動調整の指針として，音声を確認した者が，分類結果が概ね主観に合うように調整した．この際，無音クラスを別途に設け，類似度の絶対値の大きさにより無音クラスに割り振ることとした．調整された類似度ベクトルは以下の通りであった．

$$\begin{aligned} \mathbf{d}_{\mathrm{cont}} &= [1.10\ \ 0.42\ \ 0.05] \\ \mathbf{d}_{\mathrm{int}} &= [0.65\ \ 0.97\ \ 0.57] \\ \mathbf{d}_{\mathrm{imp}} &= [0.28\ \ 0.85\ \ 1.35] \end{aligned} \quad (17)$$

3.2 節で仮定を示した通り，これらの類似度ベクトルは，各信号に特徴的な帯域特徴量を強調するものとなっている．

SVM による学習は，4 セッションの会議のうち 2 セッションの正解ラベルを分類クラスとして，フーリエ分析で求めた 6 次の連続性特徴スペクトルを入力とした．SVM の学習には線形カーネルを用い，コスト関数は 100 とした．

表 1 は，正解ラベルに対するフィルタバンク分析による分類結果を，ある 1 セッションの 1 話者に対して算出した confusion matrix である．この結果から，突発音声と断続音声・無音との判別が難しいことが分かる．突発音声は類似度の大きさが小さくなり，無音に分類される場合もある．また主観分類の際，断続音声と突発音声とで判断がつきにくいものがあるため，突発音声が断続音声に分類される，という現象

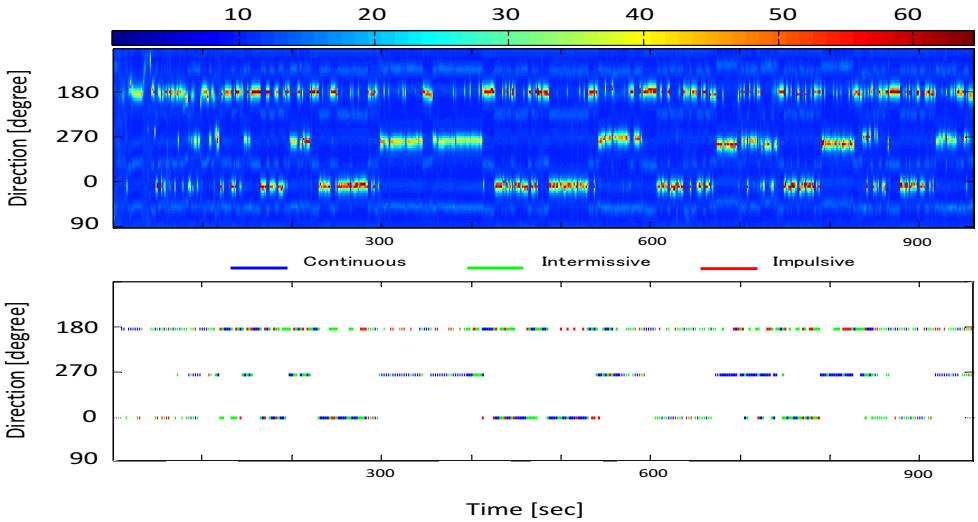

Fig. 2　カンバセーションフローによる会議全体の表現

が起きている．連続音声，断続音声については，ある精度よく分類できていると考えられるが，今後の精度向上が必要である．

表 2 は SVM による分類結果の confusion matrix である．高次の特徴量をそのまま扱っているため，連続音声と断続音声の区別，断続音声と突発音声の区別ができていない．また類似度ベクトルを用いる手法と比べ，精度が期待できる音声のクラスに差異があることが確認できる．

5　カンバセーション・フローの検討

前章での結果より，本論文で提唱するカンバセーション・フローの有効性について考察する，

図 2 は，ある 1 セッション（16 分）の会議の全体を示したものである．上段はブロック系列 MUSIC スコアのスペクトログラム表現であり，数十分程度の時間長の会議で，断続音声や突発音声の視認が難しくなる．例えば，180° 方向の話者の 300 秒付近の断続音声，0° 方向の 700 秒付近の断続音声，突発音声である．このような部分には相槌や返答が含まれている可能性があり，視認できずに聞き逃すことで，この後の会話の脈絡をつかめなくなる場合がある．

下段は本論文での類似度ベクトルに基づく音声の分類結果を 3 色で色分けして示したものである．このように各方向ごとに断続音声や突発音声を検出し，連続音声と同等の扱いで表示することにより，時間長は短いが重要な音声を聞き逃さないようにすることができると考えられる．

6　おわりに

会議録再生支援のフレームワークとして，発話状態を分類し，その情報を可視化して会議全体を表現するカンバセーション・フローを提案した．実現のための手法として，信号の方向推定スコアにフーリエ分析を用いる手法を提案し，音声クラスの分類にはパラメータ調整をしたフィルタバンク分析と，SVM の二つを用いた．

Table 1　フィルタバンク分析による分類結果の confusion matrix

正解		分類結果			
クラス	ブロック数	連続	断続	突発	無音
連続	246	69.9%	28.5%	1.6%	0.0%
断続	210	15.2%	63.8%	17.1%	3.8%
突発	112	4.7%	22.3%	33.0%	40.1%
無音	428	0.7%	0.93%	1.6%	96.7%

Table 2　SVM による分類結果の confusion matrix

正解		分類結果			
クラス	ブロック数	連続	断続	突発	無音
連続	246	56.5%	30.1%	3.7%	9.8%
断続	210	7.6%	90.5%	1.0%	1.0%
突発	112	16.1%	79.5%	1.8%	2.7%
無音	428	2.3%	7.5%	1.6%	88.6%

また，提案手法の有効性を検証する実験によって，通常の発話インデックスでは見落としやすい時間長の短い音声についても，強調して可視化することで見落しにくくなることや，会議を俯瞰して見ることで話題の展開を察しやすくなることを確認した．

参考文献

[1] S.Araki *et al.*,“Speaker indexing and speech enhancement in real meetings/conversations,” Proc. ICASSP 2008, pp.93–96, 2008.

[2] J.Pardo *et al.*,“Speaker dialization for multi-microphone meetings using only between-channel differences,” Proc. MLMI’06, pp.257–264, 2006, Springer.

[3] X.Anguera *et al.*,“Acoustic beamforming for speaker dialization of meetings,” IEEE Trans. Audio, Speech and Language Processing, vol.15, pp.2011–2022, 2007.

[4] M.Katoh *et al.*,“State estimation of meetings by information fusion using bayesian network,” Proc. Interspeech 2005, pp.113–116, 2005.