社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69（38）
2007／12／21

ポスター会話に対する発話区間検出と話者識別の検討

石塚 健太郎[†‡] 荒木 章子[†] 藤本 雅清[†] 瀬戸口 久雄[‡] 高梨 克也[*] 河原 達也[‡*]

[†]日本電信電話株式会社 NTTコミュニケーション科学基礎研究所 〒619-0237 京都府相楽郡精華町光台2-4
[‡]京都大学 大学院 情報学研究科 〒606-8501 京都市左京区吉田本町
[*]京都大学 学術情報メディアセンター 〒606-8501 京都市左京区吉田本町
E-mail: {ishizuka, shoko, masakiyo}@cslab.kecl.ntt.co.jp, {setoguchi, takanasi}@ar.media.kyoto-u.ac.jp

あらまし 会議やポスター発表などの，多人数によるインタラクションを含む場面において，「誰がいつ話したか？」を検出できれば，検索のためにインデクスを付与する場合や談話構造分析の手がかりとして有用である．この実現のためには，まず観測信号の中から何らかの音声が話されている区間を取り出し（発話区間検出），検出された音声区間について発話者を分類する必要がある（話者識別）．本稿では，マイクロホンアレイによりポスター発表を収録して得られた音声データに対し，「いつ」を捉えるために音声の周期性・非周期性の比を用いた発話区間検出技術を適用し，「誰が」を捉えるために音声信号の到来方向を用いた話者識別の手法を適用した場合について，その結果得られる話者識別性能に関し予備的な検討を行った．

キーワード 話者識別，発話区間検出，マイクロホンアレイ，多人数インタラクション

A Study on Speech Activity Detection and Speaker Diarization for the Recordings of Poster Sessions

Kentaro ISHIZUKA[†‡] Shoko ARAKI[†] Masakiyo FUJIMOTO[†] Hisao SETOGUCHI[‡]
Katsuya TAKANASHI[*] and Tatsuya KAWAHARA[‡*]

[†]NTT Communication Science Laboratories, NTT Corp., 2-4 Hikaridai, Seikacho, Sourakugun, Kyoto 619-0237 Japan
[‡]Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501 Japan
[*]Academic Center for Computing and Media Studies, Kyoto University, Sakyo-ku, Kyoto 606-8501 Japan
E-mail: {ishizuka, shoko, masakiyo}@cslab.kecl.ntt.co.jp, {setoguchi, takanasi}@ar.media.kyoto-u.ac.jp

Abstract Detecting "Who spoke when?" from multi-party interactions such as meetings and poster presentations is valuable for adding metadata to the recordings or analyzing the discourse-structures of the multi-party interactions. To realize this function, we first detect speech periods from the observed signals (speech activity detection), and then classify the speech periods by its speakers (speaker diarization). In this paper, we adopt a speech activity detection method and a speaker diarization method to the recordings of poster sessions. The speech activity method performs based on the ratios of periodic and aperiodic components of observed signals. The speaker diarization method utilizes the direction of arrival estimation of the detected speech signals obtained from a microphone array. This paper reports preliminary results obtained from these methods.

Keyword Speaker diarization, Speech activity detection, Microphone array, Multi-party interaction

1. はじめに

近年，AMI（Augmented Multi-party Interaction）[1]，CHIL（Computers In the Human Interaction Loop）[2]，NISTのRich Transcription Meeting Recognition [3]など，多人数によるインタラクションを含むデータを収録し，これらのデータを自動的に分析してインデクスを付与する研究が広く行われている．このような自動インデキシングができれば，必要な情報への効率的なアクセスが可能となり，会議録の自動作成や，要約の自動生成を行う技術の実現に繋がる．

上記の多人数インタラクション分析の研究では，多くの場合，多数のセンサを用いて複数のモダリティ情報を抽出し分析の対象とするが，本稿ではハンズフリーで収録可能な音声データのみを対象とする．これは，

簡便な設備で可能な限りの情報を抽出することを目指していることによる．

音声データから自動インデクス付与を行うための基本的な情報として「誰がいつ話したか？」を捉えられると有用である．このうち「いつ」を検出するには，観測データの中から発話のない区間を棄却し，発話のある区間を検出する必要がある．これを実現するために，発話区間検出技術（speech activity detection）を用いる[4][5]．また，「誰が」を検出するには，発話区間検出により得られた発話区間のうち，どの区間がどの発話者によるものかを分類する必要がある．これは話者識別技術（speaker diarization）を用いて実現できる[6][7]．

本稿では，複数の聞き手に対して話し手がポスターを用いて自身の研究内容について説明を行うポスター発表に伴って生じる「ポスター会話」を，多人数インタラクションを含む音声データとして収録し，収録データに対して上記の技術を併せて用いることで，「誰がいつ話したか？」を音声データから自動的に検出する手法について予備的な検討を行った結果について報告する．

2. ポスター会話の収録

2.1. ポスター会話の性質

ポスター会話は，他の多人数インタラクションを含むデータと比べ，会話のゴール／情報リソース／参加者の役割付けについて独自の性質を持つ．本節ではその性質について述べる．詳細は文献[8]を参照されたい．

今回収録したポスター会話は，話し手が1人で，聞き手が2人によるものである．話し手はポスターについて説明を行うが，聞き手は必要に応じていつでも話し手の説明をさえぎって質問やコメントを行っても良い．また，聞き手の間でのインタラクションにも制限はなく，聞き手はポスターの内容について事前に情報を持っていない．

図1のような会話のゴール／情報リソースの分類に従った会話マップに基づく分析によれば[8]，多人数インタラクションを含む会話において，ポスター会話は以下の性質を持つ．

まず，会話のゴールに関しては，雑談よりも話題のテーマや内容が決まっている一方で，チケット予約等の課題遂行型対話やディベートと比較すると話の進め方に比較的自由度がある．次に，会話で用いる情報リソースの面で考えると，必ずしも情報リソースを必要としないミーティングや，調理中の料理という動的な情報リソースを利用する料理教室の会話などの中間に位置し，ポスターという静的な情報リソースを必要とする．また，参加者の役割付けから見ると，質問等の

		Resources		
		None	Static	Dynamic
Goals	Goal-observable	Ticket Reservations	Map tasks	Cooking tutorials
	Goal-oriented	Debates	Seminars	-
	Theme-oriented	Meetings	**Poster presentations**	-
	Vague	Chattering	-	-

Fig. 1: Conversation map

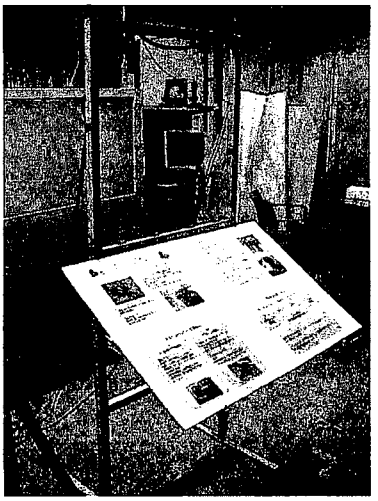

Fig. 2: Framework for presenting poster. Microphone array is set at the top of the framework (indicated by an ellipse).

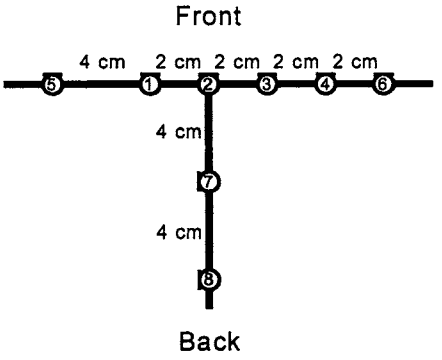

Fig. 3: Microphone array. Numbers indicate microphone indices.

時間が限られている講演と，話し手と聞き手の役割付けがないミーティングとの中間的な位置づけにあたり，参加者に話し手／聞き手の役割はあるものの，どの参加者も自由に会話の主導権を握ることができるようになっている．

2.2. 収録環境

音声データの収録は，図2に示すような，複数モダリティからの情報を収録するために作成されたポスタ

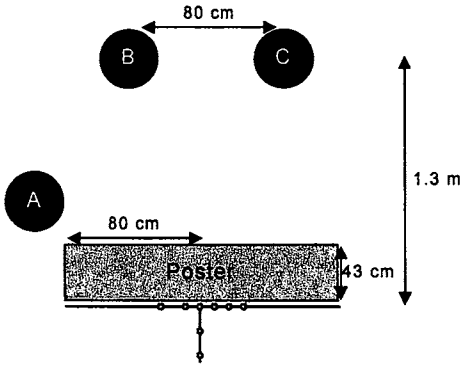

Fig. 4: Approximate speakers' position. A, B, and C indicate a poster presenter and audiences. Participants were allowed to move freely.

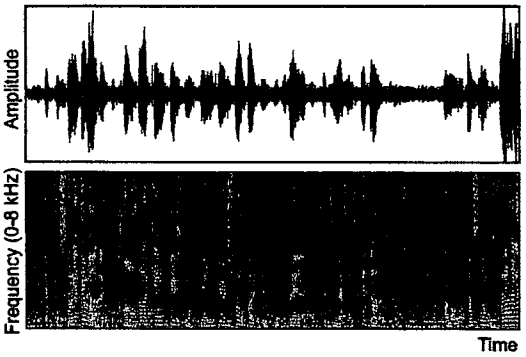

Fig. 5: Example waveform (top) and its spectrogram (bottom) for the recorded poster presentation.

一発表用のフレーム[8]の上部に無指向性マイクロホン8本から構成されるマイクロホンアレイを配置して行った．マイクロホンの配置を図3に，話し手と聞き手の収録開始時の配置を図4に示す．図4の配置は固定的なものではなく，聞き手／話し手ともに収録中は自由に移動して構わない．マイクロホンはSONY製コンデンサマイクロホン ECM-77B，マイクアンプにはYAMAHA製 HA-8，A/D変換器にはSDS製 DASBOX Model-500 を用いた．データ収録時にはサンプリング周波数 16 kHz, 量子化ビット数 24 bit で量子化を行い，3節および4節のデータ処理を行う際には量子化ビット数を 16 bit に変換して行った．

上記音声データ収録と平行して，文献[8]と同様にワイヤレスマイク，カメラ，モーションキャプチャ，アイマークレコーダによるデータの収録も行ったため，話し手／聞き手はそれらの装備を装着した状態でポスター会話を行っている．なお，1節で述べたように，本稿ではマイクロホンアレイによる収録データのみを分析対象とし，これらのデータは扱わない．

```
989.1360  989.6720  A: あるいは
990.5520  994.8560  A: 短期でしかいないんだけれども
持ってけるか持ってけないかなんか知らなかったりとか
992.1360  992.5520  B: (F_う)(F_まー)
993.5760  996.2400  B: (D_わかん) (F_んー)とりあえず
可愛いから買っちゃおうみたいな
995.1280  998.2480  A: これがチンパンジーかどうかも
よくわかんなかったりとかっていう
996.7280  996.9040  B: (F_あ)
997.8560  998.9200  B: {笑}
998.6400  998.9120  A: (F_あ)で
999.1760  999.7840  C: ほお
999.1840  1000.8160 A: ただのちっちゃい可愛い
999.3120  1002.0720 B: (L_すごいな_L)
```

Fig. 6: An example of annotations for the recorded poster presentations. Each line indicates the utterance start time (sec), end time (sec), speaker index (A, B, and C), and annotation.

以上のようにして，1回あたり15〜20分程度のポスター会話の音声データを合計8セッション収録した．収録された音声データの一例を図5に示す．室内にある計算機等の雑音が混入することにより，各マイクロホンでの信号対雑音比は 0〜6.5 dB 程度であった．

2.3. ラベルの付与

上記のようにして収録されたポスター会話8セッションのうち4セッションについて，日本語話し言葉コーパス[9]に準拠して音声を書き起こし，基本転記単位，節単位境界のラベルを付与した．書き起こしの例を図6に示す．ただし，本稿でのデータ分析の目的から，図6に含まれる言語情報は用いず，発話の開始終了時刻と話者の情報のみを用いた．以後，ラベル付与を行った4セッションについて，Session1〜4として参照する．それぞれのセッションにおいて，セッション全体のデータ長と各話者の発話時間長を表1に示す．発話の重複や発話のない時間も含まれることから，データ

Table. 1: Total data length and utterance length for each speaker. Speaker A indicates the poster presenter, and speakers B and C indicate audiences. Speakers are different for each session except for speaker A on sessions 1 and 2. Due to speech overlaps, the total data length is not equal to the summation of utterance lengths for each speaker.

	Speaker			Total
	A	B	C	
Session1	824.01	286.76	32.29	1036.89
Session2	788.98	129.45	129.01	913.42
Session3	1068.32	58.58	178.38	1149.86
Session4	1175.12	31.98	200.43	1290.66

(sec)

長と各話者の発話長の合計は一致しない．

3. 発話区間検出

3.1. 方法

本稿では，発話区間検出手法として，観測信号の周期性成分と非周期性成分の比を音響特徴として用いる手法（periodic to aperiodic component ratio based detection; PARADE）[10]を適用し，そのポスター会話における精度を検討する．本手法はモノラル録音を対象とすることから，図3のマイクロホン2に収録された音声データに対して処理を行った．PARADEの処理の概要を以下に示す．

観測信号$s(t)$が，周期性成分$s_p(t)$（ある基本周波数（F0）とその倍音成分から成る調波複合音成分）と非周期性成分$s_a(t)$（周期性成分以外の成分）との和で与えられると仮定し，それぞれの短時間周波数スペクトルを$S(n,m)$，$S_p(n,m)$，$S_a(n,m)$，短時間フレーム内のパワーを$\rho(n)$，$\rho_p(n)$，$\rho_a(n)$として，以下を仮定する．

$$|S(n,m)|^2=|S_p(n,m)|^2+|S_a(n,m)|^2 \quad (1)$$

ここでnはフレーム番号，mは離散フーリエ変換における周波数ビン（$m = 1 .. M$）を表す．$\rho(n)=(1/M)\sum_{m=1}^{M}|S(n,m)|^2$が成り立つことと上記の仮定(1)より，$\rho(n)=\rho_p(n)+\rho_a(n)$と記述できる．

次に，フレームnにおける周期性成分の基本周波数（F0）とカットオフ周波数から定まる倍音成分数をそれぞれ$f_0(n)$，$\nu(n)$とし，$f_0(n)$の整数倍の周波数ビンに含まれる非周期性成分の平均パワーが非周期性成分の全周波数ビンにおける平均パワーと等しいとみなし，以下の仮定(2)を導入する．ここで，$[kf_0(n)]$は第k倍音が含まれる周波数ビンを求める演算子を表す．

$$(1/M)\sum_{m=1}^{M}|S_a(n,m)|^2=(1/\nu(n))\sum_{k=1}^{\nu(n)}|S_a(n,[kf_0(n)])|^2 \quad (2)$$

一方，純音のパワー$\rho_s(n)$は，純音の周波数スペクトル$S_s(n,m)$と，時間長Lの左右対称な分析窓関数$g(t)$から導出される$\eta=\left(2\sum_{t=1}^{L}g(t)^2\right)\Big/\left(\sum_{t=1}^{L}g(t)\right)^2$を用いて，$\rho_s(n)=\eta|S_s(n,m)|^2$によって求めることができる．周期性成分のパワーが各倍音成分の合算で求まると仮定し，仮定(1)(2)を用いることで，観測信号$s(t)$に含まれる周期性成分と非周期性成分の推定パワー$\hat{\rho}_p(n)$，$\hat{\rho}_a(n)$を以下の(3)(4)により求めることができる（詳細な導出は文献[10]を参照）．

$$\hat{\rho}_p(n)=\eta\frac{\sum_{k=1}^{\nu(n)}|S(n,[kf_0(n)])|^2-\nu(n)\rho(n)}{1-\eta\nu(n)} \quad (3)$$

$$\hat{\rho}_a(n)=\rho(n)-\hat{\rho}_p(n) \quad (4)$$

なお，上記で必要となるF0については，本稿では自己相関法[11]を用いて推定した．

これらの推定パワーに基づき，PARADEは各フレーム内の目的音声の有無を以下のように判定する．

フレームnにおける音声／非音声区間を示す状態（1：音声，0：非音声）を変数H_nで表す．$H_n = 0$の時，非周期性成分の推定誤差$\varepsilon_a(n)$が平均0，分散$\alpha\hat{\rho}_a(n)$の正規分布に従うと仮定する．一方，$H_n = 1$の時は，周期性成分の推定誤差$\varepsilon_p(n)$が平均0，分散$\beta\hat{\rho}_p(n)$の正規分布に従うと仮定する．上記に基づき，観測信号が非音声区間／音声区間となる尤度を(5)(6)で定める（α，βは正の定数）．

$$p(\rho(n)|H_n=0)=c_1(n)\exp\left(-\left(\hat{\rho}_p(n)/\hat{\rho}_a(n)\right)^2/2\alpha^2\right) \quad (5)$$

$$p(\rho(n)|H_n=1)=c_2(n)\exp\left(-\left(\hat{\rho}_a(n)/\hat{\rho}_p(n)\right)^2/2\beta^2\right) \quad (6)$$

これらの尤度を用いて，以下の尤度比$\Lambda(n)$が閾値を上回れば音声区間として判定する．

$$\Lambda(n)=p(\rho(n)|H_n=1)/p(\rho(n)|H_n=0) \quad (7)$$

また，得られた判定結果には，ETSI ES 202 050 [5]のフレーム棄却に用いるVADと同様のHangover処理を適用した．本稿では，4節の話者識別と処理フレームレートを合わせるために，フレーム長64 ms，フレームシフト32 msで発話区間検出処理を行った．

3.2. 評価

発話区間検出の評価尺度として，NIST Rich Transcription Meeting Recognition [3]で用いられているDiarization Error Rate（DER）を用いた．DERは発話区間の誤棄却・誤受理を統合して評価する尺度であり，値が小さいほど発話区間検出性能が高いことを示す．

$$DER=\frac{\text{誤受理・誤棄却した時間長}}{\text{データの総時間長}}\times 100 \ (\%) \quad (8)$$

DERを測定する際の評価基準についても，NIST Rich Transcription Meeting Recognition [3]における評価基準に準拠した．すなわち，音声信号区間は300 ms以上の非音声信号区間で区切られるものとし，笑い声・咳などの口から発せられる非言語音は非音声区間として扱い，自動検出される発話区間の開始終了時刻は正解ラ

Table. 2: Diarization error rate (DER), missed speech time (MST), and false alarm speech time (FST) obtained from speech activity detection for each session.

	Speech Activity Detection		
	DER	MST	FST
Session1	3.9	2.1	1.8
Session2	6.6	3.1	3.5
Session3	4.0	0.6	3.4
Session4	1.3	0.8	0.5

(%)

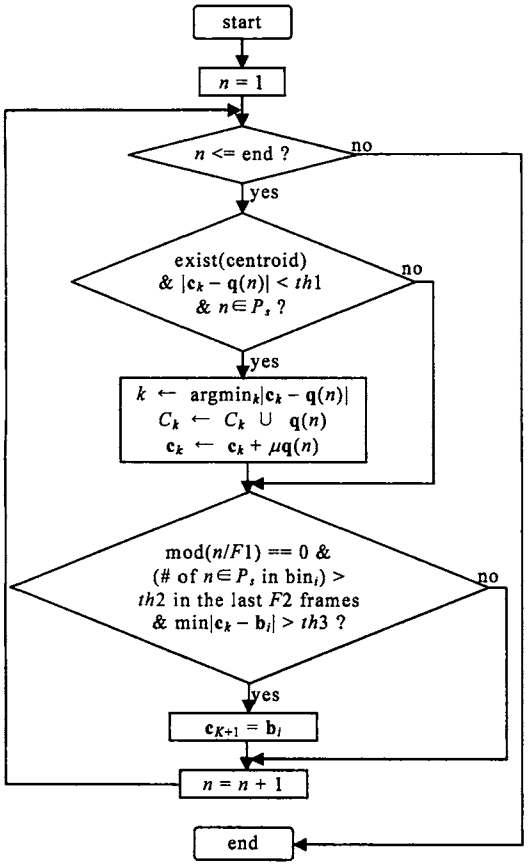

Fig. 7: Block diagram of online clustering. bin_i is the pre-defined feature range, $\mathbf{b}_i$ is a representing vector for bin_i. C_k, $\mathbf{c}_k$, $\mathbf{q}(n)$, n, and K mean cluster k, the centroid of cluster k, DOA at frame n, frame index, and number of existing clusters, respectively.

ベルに対し前後 250 ms までのずれを許容範囲とした．各セッションにおける DER，誤棄却率（Missed Speech Time; MST）および誤受理率（False-alarm Speech Time; FST）を表 2 に示す．表 2 に示される通り，DER は 1.3 ～6.6 %と低い値を示した．これは，マイクロホンに混入した主要な雑音が定常的な計算機雑音であったためと考えられる．4 節では，この発話区間検出結果に基づき話者識別の処理を行う．

4. 話者識別

4.1. 方法

本稿では，上記の発話区間検出により求まった音声区間に対して話者識別を行う手法として，観測信号の推定到来方向（direction of arrival; DOA）を用いる手法を適用した．本手法は，話者数を未知として，音声信号の推定 DOA がある一定の範囲に収まる場合には同一話者による発話とみなすことで話者のクラスタリングを行う．本稿では，DOA 推定のために，図 3 のマイクロホン 1，3，7 に収録された音声データを用いて処理を行った．処理の概要を以下に示す．

3 節で得られた発話区間を P_s とし，これを各話者の発話区間 P_k に分割する（$k = 1 .. N$; N は話者数）．そのために，P_s に含まれるフレーム n における DOA $\mathbf{q}(n)$ をクラスタリングする．DOA の推定を行うためには，まず GCC-PHAT 法（generalized cross correlation method with the phase transform）[12]を用いて各マイクロホンペア $\{j, j'\}$ について到来時間差（time difference of arrival; TDOA）$q'_{jj'}(n)$ を推定する．

$$q'_{jj'}(n) = \arg\max_{q'} \sum_f \frac{S_j(n,f)S^*_{j'}(n,f)}{\left|S_j(n,f)S^*_{j'}(n,f)\right|} e^{j2\pi fq'} \tag{9}$$

ここで f は周波数である．DOA ベクトル $\mathbf{q}(n)$ は，全てのマイクロホンペアから得られる $q'_{jj'}(n)$ からなる TDOA ベクトル $\mathbf{q}'(n)$ とマイクロホン配置の情報を表すベクトル $\mathbf{D}$ を用いて，以下で求めることができる[13]．

$$\mathbf{q}(n) = c\mathbf{D}^{-}\mathbf{q}'(n) \tag{10}$$

$$\mathbf{q}(n) \leftarrow \mathbf{q}(n)/\|\mathbf{q}(n)\| \tag{11}$$

ここで c は音速，$^-$ は一般化逆行列を表す．音源方向の方位角を $\theta(n)$，仰角を $\phi(n)$ としたとき，得られる DOA ベクトルは以下の要素を持つ．

$$\mathbf{q}(n) = [\cos\theta(n)\cos\phi(n), \sin\theta(n)\cos\phi(n), \sin\phi(n)]^T \tag{12}$$

ここから，音声信号が到来する方位角と仰角をフレームごとに推定することが可能となる．

次に，フレームごとの DOA ベクトルをクラスタリングする．話者数が未知の状態でクラスタリングを行うため，leader-follower clustering [14]によるオンラインクラスタリングを行う．このアルゴリズムのブロック図を図 7 に示す．この方法では，新たな話者が収録データに現れた場合に，新たなセントロイドを生成してクラスタリングを行う．各話者の発話区間 P_k は，

$$n \in P_k \quad \text{if} \quad \mathbf{q}(n) \in C_k \tag{13}$$

として求める．ここで C_k は k 番目のクラスタを表す．こうして得られた P_k を，話者 k による発話区間とする．図 7 で用いる定数として，本稿では，仰角（度）に関して $th1 = 15$，$th3 = 30$ を用い，フレーム数に関しては $F1 = 20$，$F2 = 500$，$th2 = 20$ を用いた．フレーム長は 64 ms，フレームシフトは 32 ms であった．

4.2. 評価

上記の手法による話者識別性能を評価するため，3.2 節と同様に，DER による評価を用いた．話者識別性能を評価する DER では，発話区間検出と異なり，誤棄却・誤検出に加えて話者の誤り（Speaker Error Time;

Table. 3: Diarization error rate (DER), missed speech time (MST), false alarm speech time (FST), and speaker error time (SET) obtained from speaker diarization for each session.

	Speaker Diarization			
	DER	MST	FST	SET
Session1	32.0	2.8	2.2	27.1
Session2	24.1	3.0	3.6	17.5
Session3	21.9	0.6	3.4	17.9
Session4	18.5	0.7	0.6	17.2

(%)

SET）についても評価に含める．

$$DER=\frac{\text{誤受理・誤棄却・話者誤りの時間長}}{\text{データの総時間長}}\times 100(\%) \quad (14)$$

評価基準は 3.2 節と同様である．各セッションにおける DER，MST，FST，SET を表 3 に示す．3 節の発話区間検出技術の結果に基づいて話者識別を行っていることから MST および FST は低いものの，SET は比較的高い値を示している．これは，本手法が話者の DOA の情報のみを利用しているため，話者の移動による影響を受けたり，立ち位置が近い聞き手 2 人の差をクラスタリングすることが困難であったりしたためである．特に図 4 の B が A に近づいたり，C が A に近づくことにより B と C の距離が近づいたりすることによって，話者の混同が起きており，Session1 と 2 では特にそれが顕著であった．また，室内／室外からの雑音は定常的であるものの方向性を持っており，発話区間であっても音声信号のパワーが小さいと方向性雑音の DOA を検出してしまい，適切な話者へ発話区間が割り当てられず，このことも SET が高くなる要因となった．

5. まとめ

本稿では，マイクロホンアレイを用いたポスター会話の収録について述べ，その収録データに対する発話区間検出と話者識別の予備的な検討を行った．

発話区間検出については，3 節で述べたとおり，信号対雑音比が低いにも関わらず，PARADE を用いることで良好な検出精度を得ることができた．しかし，複数のマイクロホンから得られる情報を統合したり，DOA のような空間的情報を利用したり[15]，他の音響特徴と併用したりする[16]ことにより，さらなる精度向上が可能と考えられる．これらについては今後検討すべき課題である．

話者識別については，DOA 推定の誤りによる話者の誤りが大きい傾向にあった．今後，DOA 推定に全てのマイクロホンの情報を用いたり，方向性雑音の影響を軽減する手段を導入したり，話者の移動の影響を受けにくい音響特徴を利用することにより，精度の向上が可能となると考えられる．

また，「誰がいつ話したか」という情報だけでなく，話者間のインタラクションに関わる情報などを音響信号から取り出すことで，収録データに対しより表現力の高いインデクス付与が可能になる．このような情報抽出についても，今後検討を行う．

文　献

[1] AMI Project : http://corpus.amiproject.org/

[2] CHIL : http://chil.server.de/

[3] NIST Rich Transcription Meeting Recognition : http://nist.gov/speech/tests/rt/

[4] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detector", IEEE Signal Processing Letters, Vol. 16, pp. 1-3, 1999.

[5] ETSI, "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; advanced front-end feature extraction algorithm; Compression algorithms", ETSI ES 202 050, v1.1.5, 2007.

[6] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 14, pp. 1557-1565, 2006.

[7] X. Anguera, C. Wooters, J. M. Pardo, and J. Hernando, "Automatic weighting for the combination of TDOA and acoustic features in speaker diarization for meetings," Proc. of ICASSP, Vol. 4, pp. 241-244, Hawaii, USA, Apr. 2007.

[8] H. Setoguchi, K. Takanashi, and T. Kawahara, "Multi-modal conversational analysis of poster presentations using multiple sensors," Proc. ACM ICMI-2007 Workshop on Tagging, Mining and Retrieval of Human Related Activity Information, pp. 44-47, Nagoya, Japan, Nov. 2007.

[9] 日本語話し言葉コーパス : http://www.kokken.go.jp/katsudo/seika/corpus/

[10] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," Proc. ISCA ITRW SAPA, pp. 65-70, Pittsburg, USA, Sep. 2006.

[11] W. Hess, *Pitch Determination of Speech Signals*, Springer-Verlag, New York, 1983.

[12] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. 24, pp. 320-327, 1976.

[13] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," Proc. ICASSP, Vol. 5, pp. 33-36, Toulouse, France, May. 2006.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification 2nd Edition*, Wiley Interscience, 2000.

[15] J. E. Rubio, K. Ishizuka, H. Sawada, S. Araki, T. Nakatani, and M. Fujimoto, "Two-microphone voice activity detection based on the homogeneity of the direction of arrival estimates," Proc. ICASSP, Vol. 4, pp. 385-388, Hawaii, USA, Apr. 2007.

[16] 藤本 雅清，石塚 健太郎，中谷 智広，"複数の音声特徴量及び信号識別処理の適応的統合に基づく音声区間検出，" 日本音響学会講演論文集，3-3-11，秋季，pp. 163-166, 2007.