論文 / 著書情報
Article / Book Information

論題(和文)	複数ピンマイクで収音された会議音声の重畳区間検出
Title(English)	Detecting Overlapped Speech in Meeting Recorded by Lapel Microphones
著者(和文)	横山諒, 那須悠, 岩野公司, 篠田浩一
Authors(English)	Ryo Yokoyama, Yu Nasu, Koji Iwano, Koichi Shinoda
出典(和文)	情報処理学会研究報告, Vol. 2012-SLP-92, No. 6,
Citation(English)	IPSJ SIG Technical Report, Vol. 2012-SLP-92, No. 6,
発行日 / Pub. date	2012, 7
権利情報 / Copyright	

Vol.2012-SLP-92 No.6
2012/7/20

複数ピンマイクで収音された会議音声の重畳区間検出

横山 諒[1,a)]　那須 悠[1,b)]　岩野 公司[2,c)]　篠田 浩一[1,d)]

概要：会議音声における音声認識や話者識別のための重畳区間検出法を提案する．会議音声は各話者のピンマイクによって収音される．抽出された二つの新しい特徴量を GMM 識別器の入力として用いる．一つは，相互スペクトル減算によって他者から伝達した音声信号を低減したパワー．もう一つは，比較的雑音の小さな環境で有効な振幅スペクトルコサイン相関である．評価実験では，4 人の話者による会議音声を用いて，未処理のパワーとパワースペクトルピアソン相関を特徴量として用いる従来手法と検出率を比較した．提案手法の検出率は 75.7%となり，従来手法の 66.8% から誤りを 26.8%削減した．

キーワード：重畳区間検出，スペクトル減算，コサイン距離

Detecting Overlapped Speech in Meeting Recorded by Lapel Microphones

Yokoyama Ryo[1,a)]　Nasu Yu[1,b)]　Iwano Koji[2,c)]　Shinoda Koichi[1,d)]

Abstract: We propose an overlapped speech detection method for speech recognition and speaker diarization of meetings, where each speaker wears a lapel microphone. Two novel features are utilized as inputs for a GMM-based detector. One is speech power after cross-channel spectral subtraction which reduces the power from the other speakers. The other is an amplitude spectral cosine correlation coefficient which effectively extracts the correlation of spectral components in a rather quiet condition. We evaluated our method using a meeting speech corpus of four persons. The accuracy of our proposed method, 75.7%, was significantly better than that of the conventional method, 66.8%, which uses raw speech power and power spectral Pearson's correlation coefficient.

Keywords: overlap speech detection, spectral subtraction, cosine distance

1. はじめに

近年，議事録の作成や会話からのデータマイニングを目的として，会議音声における音声認識 [1], [2] や話者識別 [2], [3], [4], [5], [6], [7] に関する研究が行われている．しかしながら，これらの手法は複数の人が同時に発話している重畳区間では精度が低いことが問題になっている．その解決方法としては，重畳区間を検出した後に，それらの区間を無視するか音源分離などの技術を適用して各話者の音声を扱えるようにすることが考えられる．本研究では，そのために，重畳区間検出に焦点を当てて，その精度向上を目的とする．

収録機器として，無指向性マイクやマイクロフォンアレイなどがよく用いられる．無指向性マイクはコストが低く容易に使えるが，各話者の音声信号を個別に取り出すことが困難である．マイクロフォンアレイは各話者の音声信号を分離することができるが，コストが高くキャリブレーションが困難であるという短所がある．そこで，本研究で

[1] 東京工業大学大学院情報理工学研究科計算工学専攻，東京
Department of Computer Science, Tokyo Institute of Technology, 2–12–1 Ookayama, Meguro-ku, Tokyo, 152–8552 Japan
[2] 東京都市大学環境情報学部情報メディア学科，横浜
Faculty of Environmental and Information Studies, Tokyo City University, 3–3–1 Ushikubo-nishi, Tsuzuki-ku, Yokohama, Kanagawa 224–8551 Japan
[a)] yokoyama@ks.cs.titech.ac.jp
[b)] nasu@ks.cs.titech.ac.jp
[c)] iwano@tcu.ac.jp
[d)] shinoda@cs.titech.ac.jp

は，会議音声を収録するためにピンマイクを用いる．これは装着が容易である上に，比較的低コストで各話者の音声信号を取り出すことが可能である．また，ピンマイクの代わりに会議の参加者が持っている携帯情報端末をテーブルの上に置くなどして収録することも可能である．

検出の方法としては，重畳区間/非重畳区間 GMM (Gaussian mixture model) とこれらの尤度比検定を行う GMM 識別器が広く用いられている [4], [5], [8], [9]．問題は，GMM の入力特徴量として何を用いるかというところにある．全帯域を足しあわせたパワー（全帯域パワー）は，複数のマイクで観測値が大きい場合，その区間が重畳していると判断できるため，重畳区間検出に有効である（例えば [9]）．また，クロストークによる影響に着目している研究もいくつかある [9], [10]．1 人の話者のみが発話している場合，各マイクで観測される信号は類似する傾向にあり，複数の話者が発話している場合，各観測信号は異なったものになる．例えば，Xiao ら [9] は GMM の入力特徴量として複数マイク間のパワースペクトルピアソン相関 (PPC: Power spectral Pearson's correlation coefficient) が有効であることを示した．

しかしながら，これら全帯域パワーと PPC の 2 つの特徴量にはそれぞれ問題点がある．まず，各観測信号は他の話者からの音声信号も含んでいるため，1 人の話者のみが発話している場合にも複数の話者のマイクで全帯域パワーが大きく観測される．そのため，複数の話者が発話している重畳区間であると誤って検出してしまう．次に，PPC はパワースペクトルを各フレーム近傍の各帯域におけるパワーの平均値によって正規化する．収録環境に白色性と定常性のある加法性雑音が存在している場合には，この正規化によって雑音が減算されるので有効である．しかしながら，1 人の話者が発話している場合，他の話者のマイクに伝達した信号も正規化される．そのため，各観測信号間の類似度が小さくなり，検出精度の低下に繋がる．

本研究では，会議音声における重畳区間検出のための二つの新しい特徴量を GMM 識別器の入力として用いることを提案する．一つは，相互スペクトル減算 (CCSS: Cross-channel spectral subtraction) [1] によって他者から伝達した音声信号を低減したパワー（CCSS パワー）である． CCSS は音源分離のための手法で，ピンマイクによって収録された各音声信号から他者の音声信号を取り除くのに有効である．もう一つは，振幅スペクトルコサイン相関 (ACC: Amplitude spectral cosine correlation coefficient) である．これは各帯域パワーの平均値による正規化を行わないので，1 人の話者のみが発話している区間であっても類似度を高く保つことができる．なお，収録には雑音の影響が小さいピンマイクを用いるので，正規化による雑音の減算は必要がないと考えられる．4 人の話者による会議音声を用いて精度を評価する．

本稿の構成は次の通りである．まず，第 2 節と第 3 節で CCSS パワーと ACC の二つの特徴量についてそれぞれ説明する．続いて，第 4 節で実験を示し，第 5 節にて結ぶ．

2. CCSS パワー

2.1 相互スペクトル減算 [1]

他者から伝達したパワーを低減させるために，相互スペクトル減算法 (CCSS) を導入する．これはスペクトル減算法 [11] に基づいた音源分離の手法である．

話者およびマイクの数をともに N とし，話者 i の装着しているピンマイクの ID も同じ i とする．$i = 1, 2, ..., N$ による観測信号を $X_i(f,t)$ とする．伝達系が線形であることを仮定し，話者の音声以外の雑音を考えないものとすると，観測信号は

$$X_i(f,t) = \sum_{j=1}^{N} G_{ij}(f,t) S_j(f,t) \tag{1}$$

とモデル化される．ここで $S_j(f,t)$ は話者 $j = 1, 2, ..., N$ の音声，$G_{ij}(f,t)$ は話者 j からマイク i への伝達関数を表す．

マイク j で観測される，対応する話者 j の音声を

$$Y_j(f,t) = G_{jj}(f,t) S_j(f,t) \tag{2}$$

とおき，伝達関数を

$$H_{ij}(f,t) = \frac{G_{ij}(f,t)}{G_{jj}(f,t)} \tag{3}$$

によって置き換えると，観測信号は

$$X_i(f,t) = Y_i(f,t) + \sum_{j \neq i} H_{ij}(f,t) Y_j(f,t) \tag{4}$$

と表される．

観測信号のパワースペクトルは

$$\begin{aligned} &|X_i(f,t)|^2 \\ &= \left| Y_i(f,t) + \sum_{j \neq i} H_{ij}(f,t) Y_j(f,t) \right|^2 \\ &= |Y_i(f,t)|^2 + \sum_{j \neq i} |H_{ij}(f,t) Y_j(f,t)|^2 \\ &\quad + \sum_{k=1}^{N} \sum_{j \neq i} |H_{ik}(f,t) Y_k(f,t) H_{ij}(f,t) Y_j(f,t)| \cos\theta_{kj,i} \end{aligned} \tag{5}$$

となる．$\theta_{kj,i}$ はマイク i で観測される話者 k の音声と話者 j の音声の位相差である．

ここで，各時間周波数において異なる話者の音声の位相は無相関であることを仮定できるため，$\cos\theta_{kj,i}$ の期待値

は 0 である，また音声信号の近似的なスパース性，すなわち $j \neq k$ の音声信号に対し

$$S_j(f,t)S_k(f,t) \simeq 0 \quad (j \neq k) \tag{6}$$

が成り立つとすると，式 (5) の第 3 項は十分小さく無視できる．従って，目的信号は

$$|\hat{Y}_i(f,t)|^2 = |X_i(f,t)|^2 - \sum_{j \neq i} |\hat{H}_{ij}(f,t)|^2 |\hat{Y}_j(f,t)|^2 \tag{7}$$

と推定できる．

続いて，伝達関数を N 人の話者のうち 1 人だけが発話しているフレームを用いて推定する．話者 j のみが発話しているフレームでは，各時間周波数でその話者の着けているマイク j で観測されるパワーが最大となることが期待される．従って，

$$|\hat{Y}_i(f,t)|^2 = \max\left(|X_i(f,t)|^2 - \sum_{j \neq i} |X_j(f,t)|^2, 0 \right) \tag{8}$$

として，適当な閾値 $T_{j1}(t), T_{k2}(t)$ を用いて

$$\frac{1}{|F_1|} \sum_{f \in F_1} |\hat{Y}_j(f,t)|^2 > T_{j1}(t) \tag{9}$$

$$\frac{1}{|F_1|} \sum_{f \in F_1} |\hat{Y}_k(f,t)|^2 < T_{k2}(t), \quad \forall k \neq j \tag{10}$$

が共に成り立つフレーム t を話者 j のみが発話しているフレームであると推定する．ここで F_1 は周波数帯域である．

話者 j のみが発話しているとき，各マイクの観測信号のパワースペクトルは式 (4) より

$$X_i(f,t) = \begin{cases} Y_j(f,t) & \text{if } i = j \\ H_{ij}(f,t)Y_j(f,t) & \text{otherwise} \end{cases} \tag{11}$$

であるため，これらの比が $|H_{ij}(f,t)|^2$ の推定値となる．

伝達関数は時刻に対して連続的に変化すると考えられるので，逐次更新を行って推定する．適当な初期値 $|\hat{H}_{ij}(f,0)|^2$ を与え，忘却関数を $\rho_h \in [0,1]$ として，話者 j のみが発話しているフレームで

$$|\hat{H}_{ij}(f,t)|^2 = \rho_h |\hat{H}_{ij}(f,t-1)|^2 + (1-\rho_h)\frac{|X_i(f,t)|^2}{|X_j(f,t)|^2} \tag{12}$$

のように更新する．

分離信号は，推定した伝達関数を用いて反復操作により推定する．初期値を $|\hat{Y}_i^{(0)}(f,t)|^2 = |X_i(f,t)|^2$ とし，適当な回数だけ

$$|\hat{Y}_i^{(n)}(f,t)|^2 = |\hat{X}_i(f,t)|^2 - \alpha_n \sum_{j \neq i} |\hat{H}_{ij}(f,t)|^2 |\hat{Y}_j^{(n-1)}(f,t)|^2 \tag{13}$$

とする更新を繰り返す．α は各反復の減算係数である．

1 回目の減算では，強調したい目的話者の音声による成分も引いてしまうため，歪みが生じる．反復を行うことにより目的話者の音声を残し，それ以外の成分を抑圧することができる．

2.2 パワーの計算

一般的に従来手法 [9] で用いられているような全帯域パワー (P) は，観測信号 $X_i(f,t)$ の各時間周波数における振幅を二乗し，全帯域 F_2 に関して和をとることで

$$\mathrm{P}_i(t) = \sum_{f \in F_2} |X_i(f,t)|^2 \tag{14}$$

と計算される．しかしながら，各観測信号は他の話者からの音声信号も含んでいるため，1 人の話者のみが発話している場合にも複数の話者のマイクで全帯域パワーが大きく観測される．そのため，非重畳区間と重畳区間の区別が曖昧になることがある．

そこで提案手法では，CCSS を施して式 (13) から推定された各話者の音声信号 $\hat{Y}_i^{(n)}(f,t)$ をパワーの計算に用いる．これを CCSS パワーと称し，

$$\mathrm{CCSS\_P}_i(t) = \sum_{f \in F_2} |\hat{Y}_i^{(n)}(f,t)|^2 \tag{15}$$

のようにして算出する．CCSS パワーは他者から伝達した音声信号を低減しているので，1 人の話者のみが発話している場合には 1 つのマイクからのみ CCSS パワーが大きく観測される．従って，非重畳区間と重畳区間の区別が明瞭になると期待できる．

3. スペクトル類似度

3.1 条件による類似度の違い

クロストークによる影響を重畳区間検出に利用する．会議において，1 人の話者のみが発話している場合，各マイクで観測される信号の類似度は大きくなり，複数の話者が発話している場合，各観測信号の類似度は小さくなる．ここでは簡単のため，2 人の話者による会議という状況に限定して説明する．

話者を i, j とし，話者 i のみが発話しているとする．各マイク i, j による観測信号のパワースペクトルは式 (7) より

$$|X_i(f,t)|^2 = |\hat{Y}_i(f,t)|^2 \quad (16)$$
$$|X_j(f,t)|^2 = |\hat{H}_{ji}(f,t)|^2|\hat{Y}_i(f,t)|^2 \quad (17)$$

となる．従って各観測信号間には

$$|X_j(f,t)|^2 = |\hat{H}_{ji}(f,t)|^2|X_i(f,t)|^2 \quad (18)$$

という関係があり，類似度が大きくなる．

次に，話者 i,j の 2 人がともに発話しているとする．この場合に各観測信号のパワースペクトルは

$$|X_i(f,t)|^2 = |\hat{Y}_i(f,t)|^2 + |\hat{H}_{ij}(f,t)|^2|\hat{Y}_j(f,t)|^2 \quad (19)$$
$$|X_j(f,t)|^2 = |\hat{H}_{ji}(f,t)|^2|\hat{Y}_i(f,t)|^2 + |\hat{Y}_j(f,t)|^2 \quad (20)$$

となる．従って各観測信号間の類似度は小さくなる．

3.2 パワースペクトルピアソン相関

従来手法 [9] では，マイク i,j で観測されたパワースペクトル間のパワースペクトルピアソン相関 (PPC) がスペクトル類似度の指標として用いられている．PPC は

$$\mathrm{PPC}_{i,j}(t) = \frac{\big(\boldsymbol{P}_i(t) - \bar{\boldsymbol{P}}_i(t)\big)\cdot\big(\boldsymbol{P}_j(t) - \bar{\boldsymbol{P}}_j(t)\big)}{\big\|\boldsymbol{P}_i(t) - \bar{\boldsymbol{P}}_i(t)\big\|\,\big\|\boldsymbol{P}_j(t) - \bar{\boldsymbol{P}}_j(t)\big\|} \quad (21)$$

と定義される．ここで，$\boldsymbol{P}_i(t)$ は周波数領域 $f \in F_3$, 時間領域 $t-T \le \tau \le t+T$ で $|X_i(f,\tau)|^2$ を並べた $|F_3|\times(2T+1)$ 次元ベクトル，$\bar{\boldsymbol{P}}_i(t)$ は窓幅 $2T+1$ フレームに関してすべての周波数帯域 $|F_3|$ におけるパワーの平均値である．

PPC は二つの信号間のスペクトル類似度を表しており，1 人の話者のみが発話している区間では大きくなり，複数の話者が発話している区間では小さくなる．平均値ベクトル $\bar{\boldsymbol{P}}_i(t)$ によって正規化することで，加法性雑音が常在している場合はその雑音を取り除くことができるので有効だといえる．しかしながら，音声信号もまた正規化されるので，単一話者が発話している区間では正規化によって PPC が小さくなる傾向があるという問題がある．

3.3 振幅スペクトルコサイン相関

本研究では，マイク i,j で観測された振幅スペクトル間の振幅スペクトルコサイン相関 (ACC) をスペクトル類似度の指標として用いる．ACC は式 (21) から正規化をする工程を除去して

$$\mathrm{ACC}_{i,j}(t) = \frac{\boldsymbol{A}_i(t)\cdot\boldsymbol{A}_j(t)}{\|\boldsymbol{A}_i(t)\|\,\|\boldsymbol{A}_j(t)\|} \quad (22)$$

と定義する．ここで $\boldsymbol{A}_i(t)$ は周波数領域 $f \in F_3$, 時間領域 $t-T \le \tau \le t+T$ で $|X_i(f,\tau)|$ を並べた $|F_3|\times(2T+1)$ 次元ベクトルである．

パワースペクトルにおける相関は，二乗によって値の大きい帯域の影響が強調され，相対的に他の帯域の情報が失われることがある．そこで本研究では，振幅スペクトルを用いた．また，雑音環境下では PPC の方が精度が高いといえるが，ピンマイクで収録された会議音声のような比較的雑音の小さい環境下では，正規化してない ACC の方が精度が高いと期待できる．

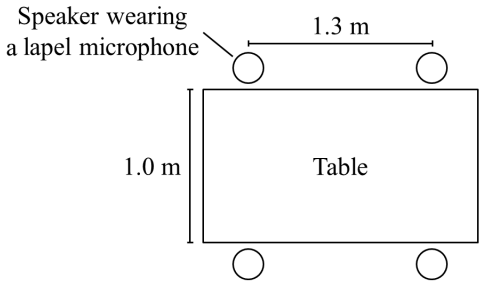

図 1 各話者の座席位置.
Fig. 1 Position of speakers in sit-down meeting.

表 1 データベースの詳細.
Table 1 Training and test dataset.

	Length	W_n	W_o
Train	9.7 min	70%	30%
Test	9.7 min	68%	32%

4. 評価実験

4.1 実験条件

評価データとして，男性話者 3 人，女性話者 1 人による 19 分間の会議音声を収録した．収録音声の前半部分を学習に，後半部分を評価に用いた．各話者の座席は図 1 の位置で，会議中の話者の移動はなかったが，姿勢は自由に変えることができる状況であった．使用したマイクは単一指向性ピンマイクで，各話者の胸元にクリップで装着した．音声区間の正解ラベルは人手によって作成されたものを用い，重畳区間ラベル (W_o) と非重畳区間ラベル (W_n) は各フレーム（25 ms 毎）に与えられている．学習データ，評価データに対する各ラベルの割合は表 1 の通りである．

収録の標本化周波数を 16 kHz とし，STFT はフレーム長 50 ms フレームシフト 25 ms のハミング窓により行った．また，CCSS における 1 人だけが発話しているフレームの推定では，閾値を

$$T_{j1}(t) = \frac{2}{|F_1|}\sum_{f\in F_1}|\hat{N}_j(f,t)|^2 \quad (23)$$
$$T_{k2}(t) = \frac{1}{|F_1|}\sum_{f\in F_1}|\hat{N}_k(f,t)|^2 \quad (24)$$

とした．ここで，$F_1 = [50, 4000]$ Hz，$|\hat{N}_i(f,t)|^2$ は推定した背景雑音のパワースペクトルである．伝達関数の推定には STFT による各周波数成分をそのまま扱うと誤差が大きいため，実験では全周波数帯域 $|F_2| = 400$ を 10 分割して，各分割帯域毎の平均値を用いて計算した．伝達関数の更新

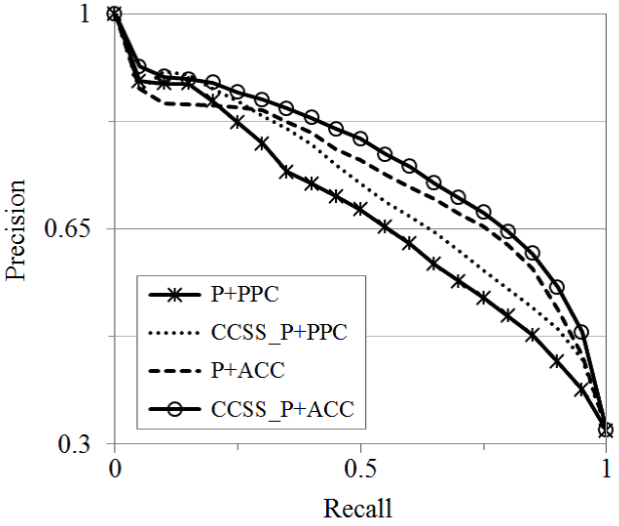

図 2 会議音声における重畳区間検出の Recall-Precison カーブ.
Fig. 2 Recall-Precison curve of overlapped detection in meeting.

は $\rho_h = 0.98$, $|\hat{H}_{ij}(f,0)|^2 = 0.2$ として行った．式 (13) による分離信号の推定には $\alpha_1 = 1, \alpha_2 = 4$ として計算した．スペクトル類似度の計算では $T = 10$ とし，全周波数領域の半分 $F_3 = [50, 4000]$ の 200 点を用いた．GMM は対角共分散行列を持ち，混合分布数は 8 とした．

GMM の入力特徴量として，従来手法 [9] で用いられている全帯域パワー (P) と パワースペクトルピアソン相関 (PPC)，提案手法で用いる CCSS パワー (CCSS_P) と振幅スペクトルコサイン相関 (ACC) を各フレームから抽出した．P と CCSS_P は 4 つのピンマイクの観測信号から抽出できるので次元は 4，また，PPC と ACC の次元は 4 人の話者の組合せから 6 となる．ここで，従来手法 [9] を P+PPC，提案手法を CCSS_P+ACC と表記する．

重畳区間の判定は，各フレーム s の W_o に対する尤度 $P(s|W_\mathrm{o})$ 及び W_n に対する尤度 $P(s|W_\mathrm{n})$ の対数尤度比

$$\begin{aligned}\Lambda(s) &= \ln \frac{P(s|W_\mathrm{o})}{P(s|W_\mathrm{n})} \\ &= \ln[P(s|W_\mathrm{o})] - \ln[P(s|W_\mathrm{n})] \end{aligned} \quad (25)$$

をある閾値 T_border を用いて

$$s = \begin{cases} \text{Overlapped speech}, & \Lambda(s) > T_\mathrm{border} \\ \text{Non-overlapped speech}, & \text{otherwise} \end{cases} \quad (26)$$

として行った．

検出精度を評価するために，式 (26) における閾値 T_border を変化させて再現率と適合率を求め，Recall-Precision カーブを作成した．そして，そこから再現率と適合率の両方を考慮することのできる平均適合率 (AP: Average precision) [12] を求め，評価基準として用いた．

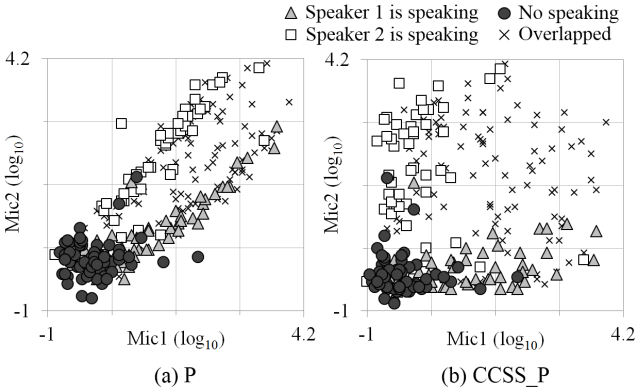

図 3 重畳区間による結果の散布図．横軸がマイク 1 による観測信号のパワー，縦軸がマイク 2 による観測信号のパワー.
Fig. 3 A scatter diagram of frame labels obtained by OSD. The horizontal axis is the power recorded by the 1st microphone, and the vertical axis is the power recorded by the 2nd microphone.

表 2 各スペクトル類似度の AP (%).
Table 2 AP (%) of each spectral similarity.

	PPC	ACC	APC	PCC
AP	44.5	50.6	49.6	44.7

4.2 実験結果

検出結果を図 2 に示す．提案手法 CCSS_P+ACC の AP は 75.7%となり，従来手法 P+PPC の 66.8%から誤りを 26.8%削減した．

データセットから特定の 2 人の話者のみが発話している区間だけを用いて，検出精度を分析した．従来法 P と提案法 CCSS_P の精度比較を図 3 に示す．CCSS_P の重畳区間領域と非重畳区間領域は P に比べると明確に区分けされてる．PPC，ACC，振幅スペクトルピアソン相関 (APC: Amplitude spectral Pearson's correlation coefficient)，パワースペクトルコサイン相関 (PCC :Power spectral cosine correlation coefficient) の検出精度を比較した結果を表 2 に示す．ACC の AP が 50.6%となり最も高い精度を示した．また，ACC と PPC の重畳区間と非重畳区間におけるヒストグラム図 4 から ACC の分布がより明確に分割されている．

CCSS_P と ACC を組合せることがどのように精度向上に繋がっているのかを分析した．Recall = 0.7 に固定した場合の，重畳区間として誤検出をした非重畳区間フレームにおけるパワーヒストグラムを図 5 に示す．CCSS_P は比較的パワーの大きなフレームで誤検出が多い．一方，ACC は比較的パワーの小さなフレームで誤検出が多いことが示さてれいる．従って，二つの特徴量を組合せることによって精度向上に繋がったと考える．

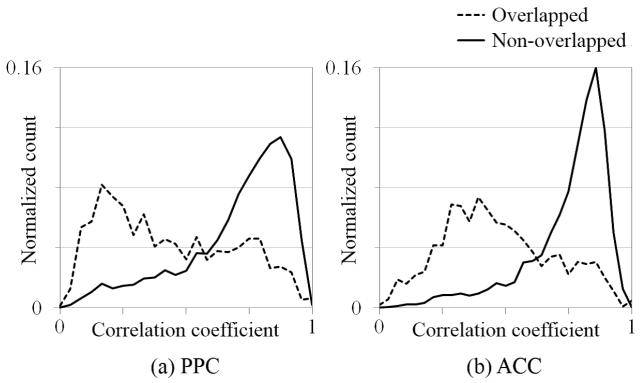

図 4 重畳フレームと非重畳フレームにおける相関係数のヒストグラム．

Fig. 4 Correlation coefficient histogram of the overlapped and non-overlapped frames.

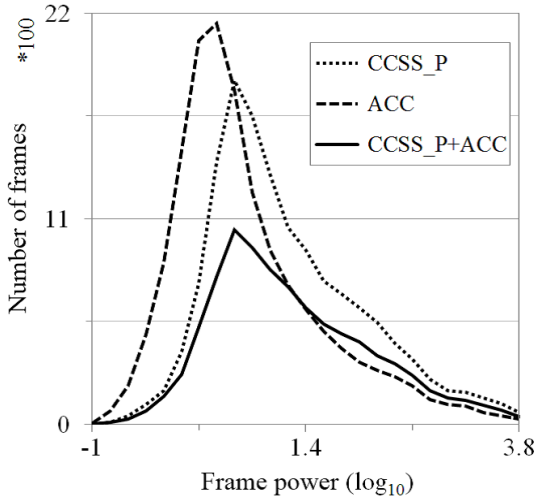

図 5 誤検出フレームのパワーヒストグラム．

Fig. 5 Power histogram of the misdetected frames.

5. おわりに

本研究では，会議音声における重畳区間検出の精度向上を目的として CCSS_P と ACC を GMM の入力特徴量として用いることを提案した．評価実験では，P と PPC を GMM の入力特徴量として用いた従来手法と検出精度を AP により比較した．提案手法の AP は 75.7%となり，従来手法の 66.8%から誤りを 26.8%削減した．

これらの精度向上にもかかわらず，誤検出区間が未だに存在している．そこで，エントロピーなど他の特徴量を組み合わせることが更なる精度向上に必要である．また，本研究では人手によるラベルを用いたが，実際はそのようなラベルを入手することは困難であり，検出コスト削減のためにも教師なし学習の導入が必要である．

参考文献

[1] Y. Nasu, K. Shinoda, and S. Furui, “Cross-channel spectral subtraction for meeting speech recognition,” in *Proc. ICASSP*, 2011, pp. 4812–4815.
[2] A. Stolcke, G. Friedland, and D. Imseng, “Leveraging speaker diarization for meeting recognition from distant microphones,” in *Proc. ICASSP*, 2010, pp. 4390–4393.
[3] F. Valente, D. Vijayasenan, and P. Motlicek, “Speaker diarization of meetings based on speaker role n-gram models,” in *Proc. ICASSP*, 2011, pp. 4416–4419.
[4] O. Ben-Harush, H. Guterman, and I. Lapidot, “Frame level entropy based overlapped speech detection as a preprocessing stage for speaker diarization,” in *IEEE International Workshop on Machine Learning for Signal Prcessing*, 2009, pp. 1–6.
[5] K. Boakye, O. Vinyals, and G. Friedland, “Improved overlapped speech handling for speaker diarization,” in *Proc. INTERSPEECH*, 2011, pp. 941–944.
[6] D. Vijayasenan, and F. Valente, “Speaker diarization of meetings based on large tdoa feature vectors,” in *Proc. ICASSP*, 2012, pp. 4173–4176.
[7] E. Zwyssig, S. Renals, and M. Lincoln, “On the effect of snr and superdirective beamforming in speaker diarisation in meetings,” in *Proc. ICASSP*, 2012, pp. 4177–4180.
[8] H. Sun and B. Ma, “Study of overlapped speech detection for NIST SRE summed channel speaker recognition,” in *Proc. INTERSPEECH*, 2011, pp. 2345–2348.
[9] B. Xiao, P.K. Ghosh, P. Georgiou, and S.S. Narayanan, “Overlapped speech detection using long-term spectro-temporal similarity in stereo recording,” in *Proc. ICASSP*, 2011, pp. 5216–5219.
[10] S.N. Wrigley, G.J. Brown, V. Wan, and S. Renals, “Speech and crosstalk detection in multichannel audio,” in *IEEE Transactions on Speech and Audio Processing*, 2004, vol. 13, no. 1, pp. 84–91
[11] S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” in *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1979, vol. 27, no. 2, pp. 113–120
[12] M. Zhu, “Recall, Precision and Average Precision,” *Technical Report 09, Department of Statistics and Actuarial Science, University of Waterloo*, 2004.