# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | 相互スペクトル減算と振幅スペクトル相関を用いた 会議音声の重畳区間検出 |
| Title(English) | Overlapped speech detection for meeting using cross channel spectral subtraction and similarity of ampli- tude spectrum. |
| 著者(和文) | 横山諒, 那須悠, 篠田浩一, 岩野公司 |
| Authors(English) | Ryo Yokoyama, Yu Nasu, Koichi Shinoda, Koji Iwano |
| 出典(和文) | 日本音響学会2012年春季研究発表会講演論文集, Vol. , No. , pp. 13-14 |
| Citation(English) | , Vol. , No. , pp. 13-14 |
| 発行日 / Pub. date | 2012, 3 |

# 相互スペクトル減算と振幅スペクトル相関を用いた会議音声の重畳区間検出*

☆横山諒, 那須悠, 篠田浩一 (東工大), 岩野公司 (都市大)

## 1 はじめに

会議音声には複数の話者が同時に発話している区間 (音声重畳区間) が存在する. 会議音声の認識や話者識別においては, そのような音声重畳区間を, 単一話者が発話している区間 (非重畳区間) と区別することが必要である.

重畳区間検出の手法としては, GMM による重畳区間と非重畳区間のモデル化が広く用いられている. 入力特徴量として, スペクトルパワー領域で定義されたエントロピーの値 [1] や, 観測信号間のパワースペクトルピアソン相関値とパワー [2] などを用いた研究がなされている.

本稿では, GMM の入力特徴量に観測信号間のスペクトル相関値とパワーを用いた手法に基づいて, 各特徴量を改善した手法を提案する. パワーとしては, 各観測信号に対する相互スペクトル減算 (CCSS) [3] によって得られるパワー (CCSS パワー) を用いる. スペクトル相関値としては, 観測信号間の振幅スペクトルコサイン相関値を用いる.

## 2 相互スペクトル減算 (CCSS) 法 [3]

話者およびマイクの数を $N$ とする. STFT による周波数領域における表現で, マイク $i = 1, 2, ..., N$ による観測信号を $X_i(f,t)$ とし, マイク $i$ で観測される話者 $i$ の音声を $Y_i(f,t)$ とする. CCSS 法により $Y_i(f,t)$ のパワースペクトルは

$$|\hat{Y}_i(f,t)|^2 = \max\left(|X_i(f,t)|^2 - \sum_{j\neq i}|X_j(f,t)|^2, 0\right) \tag{1}$$

と近似される. したがって, マイク $i$ による観測信号から, 話者 $i$ によるパワーのみを抽出したパワー (CCSS パワー) $\gamma_i(m)$ は

$$\gamma_i(m) = \sum_{f\in F}|\hat{Y}_i(f,m)|^2 \tag{2}$$

となる. CCSS パワーはパワーに比べ, クロストークによる影響が軽減されていると考えられる.

## 3 振幅スペクトル相関

話者 $j$ からマイク $i$ への減衰係数を $\alpha_{i,j}$ $(0 < \alpha_{i,j} < 1)$ とする. 話者が 2 人の場合, 各マイクによる観測信号の振幅スペクトルは

$$\begin{aligned}|\hat{X}_1(f,t)| &= |Y_1(f,t)| + \alpha_{1,2}|Y_2(f,t)| \\ |\hat{X}_2(f,t)| &= \alpha_{2,1}|Y_1(f,t)| + |Y_2(f,t)|\end{aligned} \tag{3}$$

と近似できる. したがって, 片方の話者のみが発話している場合, ベクトル $\hat{\boldsymbol{X}}_\mathbf{1}$ と $\hat{\boldsymbol{X}}_\mathbf{2}$ の向きは一致する.

ベクトル振幅スペクトルを

$$A_{x_i}(f,t) = |X_i(f,t)| \tag{4}$$

とし, コサイン類似度を用いて, 時間 $m - P \leq t \leq m + P$ における観測信号間のスペクトル相関値を表すと

$$\varphi_{i,j}(m) = \frac{\sum_{f\in F,t} A_{x_i}(f,t)A_{x_j}(f,t)}{\sqrt{\sum_{f\in F,t}\left(A_{x_i}(f,t)\right)^2}\sqrt{\sum_{f\in F,t}\left(A_{x_j}(f,t)\right)^2}}$$
$$i, j \in \{1, 2, ..., N\} \tag{5}$$

となる.

## 4 提案手法

GMM によって重畳区間と非重畳区間のモデル化を行った. 入力特徴量として式 (2) と 式 (5) で示した $\gamma_i(m)$ と $\varphi_{i,j}(m)$ を用いた.

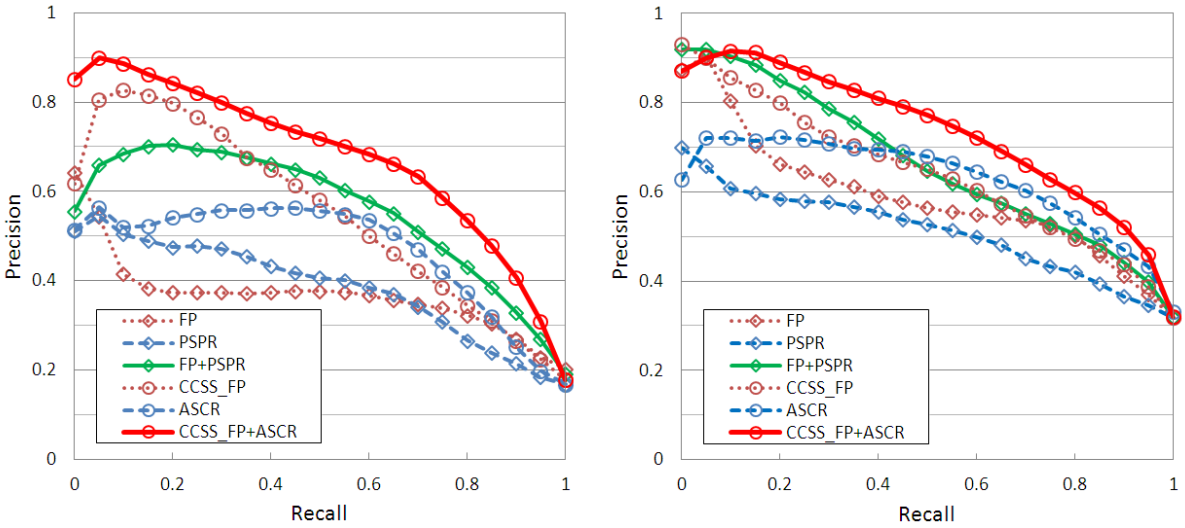
## 5 評価実験

### 5.1 実験条件

#### 5.1.1 データベース

データとして, 男性話者 3 人, 女性話者 1 人による 20 分間の会議音声と, その会議音声において女性話者とその向かい合う男性話者のみが発話している区間を切り出した 15 分間の対談音声を用いた. 音声重畳区間の割合はそれぞれ 30.9%, 16.5% であった. それぞれ前半部分を学習セット, 後半部分を評価セットとした. 会議中の話者移動はなく, 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

* Overlapped speech detection for meeting using cross channel spectral subtraction and similarity of amplitude spectrum. by Ryo Yokoyama, Yu Nasu, Koichi Shinoda (Tokyo Institute of Technology), and Koji Iwano (Tokyo City University)




(a) Conversation (b) Meeting

Fig. 1 Recall-Precision of overlapped speech detection in conversation and meeting.

装着してある. 学習に使用した重畳区間/非重畳区間の正解ラベルは人手によって作成されたものであり, 話者の笑い声や咳も音声区間に含めている.

収録音声のサンプリングレートは 16 kHz である. STFT はフレーム長 320 点, フレームシフト 160 点, Hamming 窓により行い, 周波数は 0-4 kHz のうち均等に 80 点とった. $\varphi_{i,j}(m)$ のパラメータ $P$ は 25 フレーム (10 ms 毎) とした.

#### 5.1.2 音響特徴量

GMM の入力特徴量として, パワー (FP), CCSS パワー (CCSS_FP), パワースペクトルピアソン相関 (PSPR), 振幅スペクトルコサイン相関 (ASCR) を用意した. 従来手法は FP+PSPR, 提案手法は CCSS_FP+ASCR である. GMM の学習には HTK [4] を用い, 8 混合で行った.

#### 5.1.3 評価方法

対談音声と会議音声における重畳区間検出精度を比較する. 比較には, 横軸に再現率, 縦軸に適合率をとった Recall-Precision 曲線を用いた.

### 5.2 実験結果

対談音声を用いた実験結果を Fig. 1(a) に示す. 入力特徴量にパワーのみを用いた場合を比較すると, FP より CCSS_FP のほうが良い結果を示した. 同様に, スペクトル相関を比較すると, PSPR より ASCR のほうが良い結果を示した. また, 従来手法である FP+PSPR と提案手法の CCSS_FP+ASCR を比較すると, 提案手法に大きな改善がみられた.

また, 会議音声を用いた実験結果 Fig. 1(b) においても, 従来手法 FP+PSPR に比べ, 提案手法 CCSS_FP+ASCR の検出精度に改善がみられた.

## 6 まとめ

会議音声の重畳区間検出を目的とした特徴量を提案した. 対談音声, 会議音声を用いた実験について, パワー, スペクトル相関それぞれの性能を比較しても, 両方の実験で有効性がみられた. 提案手法である, CCSS_FP と ASCR を組み合わせた CCSS_FP+ASCR についても, 従来法の FP+PSPR と比べると大きな改善が見られた.

本稿では, GMM の作成に人手による重畳区間/非重畳区間ラベルを用いたが, 実用環境においては, そのようなラベルの入手は困難である. 教師なし学習による重畳区間検出を今後の研究課題としたい.

## 参考文献

[1] O. Ben-Harush *et al.*, "Frame level entropy based overlapped speech detection as a pre-processing stage for speaker diarization," IEEE Machine Learning for Signal Prcessing, pp. 1-6, 2009

[2] B. Xiao *et al.*, "Overlapped speesh detection using long-term spectro-temporal similarity in stereo recording," ICASSP, 5216-5219, 2011.

[3] Y. Nasu *et al.*, "Cross-channel spectral subtraction fro meeting speech recognition," ICASSP, 4812-4815, 2011.

[4] Hidden Markov Model Tool Kit (HTK), `http://htk.eng.cam.ac.uk/`.