論文 / 著書情報
Article / Book Information

論題(和文)	会議音声認識のためのスペクトル減算に基づく音源分離
Title(English)	
著者(和文)	那須 悠, 篠田 浩一, 古井 貞熙
Authors(English)	Yu Nasu, Koichi Shinoda, SADAOKI FURUI
出典(和文)	日本音響学会2010年秋季講演論文集, Vol. , No. 3-10-13, pp. 627-630
Citation(English)	, Vol. , No. 3-10-13, pp. 627-630
発行日 / Pub. date	2010, 9

会議音声認識のためのスペクトル減算に基づく音源分離*

☆那須悠, 篠田浩一, 古井貞熙 (東工大)

1 はじめに

会議音声の認識は，議事録の作成や会話からのデータマイニングなどの目的のために有用な技術である．会議ではしばしば複数の話者が同時に発話する．そのため，精度の高い認識は容易ではない [1]．この問題を解決するため，音声認識の前処理としての音源分離手法の研究が行われている．

音源分離の手法としては，複数のマイクを用いて観測した多チャンネルの信号に対して行う方法で高い効果が示されており，独立成分分析によるもの [2, 3]，音声のスパース性に基づくバイナリマスクによるもの [4, 5] などが提案されている．しかし，独立成分分析による手法では高次統計量の計算を行うことから，必要とする演算量が大きい．バイナリマスクによる手法は，音源信号のスパース性が十分高くない場合や，マスクの推定に誤りが生じた場合には分離信号の歪みが大きくなる．また，従来研究では複数のマイクを近距離に固めて配置し，話者の位置も固定されていることを仮定することが多いが，実際の会議音声認識に利用するには課題がある．

本稿では，会議音声の認識を目的とし，スペクトル減算に基づく音源分離手法を提案する．音声入力は各話者が胸元に装着するピンマイクによって行う．提案手法は伝達特性を逐次的に推定し，話者の姿勢などによる変化に対応する．独立成分分析による手法に比べ演算量が小さく，実時間処理が可能である．またバイナリマスクによる手法より頑健性が高い．

2 スペクトル減算法

単一のマイクを用いて目的音の強調を行う手法の一つとして，スペクトル減算法 [6] が広く用いられている．スペクトル減算法は，短時間フーリエ変換 (STFT) による周波数領域における表現で，観測信号 $X(f,t)$ を目的信号 $S(f,t)$ と加法性雑音 $N(f,t)$ の和としてそのパワーを

$$|X(f,t)|^2 \approx |S(f,t)|^2 + |N(f,t)|^2 \tag{1}$$

と近似し，雑音パワーの推定値 $|\hat{N}(f,t)|^2$ を減算することによって目的信号の推定パワー

$$|\hat{S}(f,t)|^2 = |X(f,t)|^2 - \alpha|\hat{N}(f,t)|^2 \tag{2}$$

を得る．α は減算係数である．通常 $|\hat{N}(f,t)|^2$ には t によらない推定値が用いられる．

3 スペクトル減算に基づく音源分離

3.1 アルゴリズム

話者およびマイクの数を N とし，マイク $i = 1, 2, \ldots, N$ 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

$$X_i(f,t) = \sum_{j=1}^{N} G_{ij}(f,t) S_j(f,t) \tag{3}$$

とモデル化される．ここで，$S_j(f,t)$ は話者 $j = 1, 2, \ldots, N$ 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．従来研究の多くでは伝達関数を時不変としているが，実環境では話者の動きにより変化しうるため，ここでは時刻の関数とする．

音源分離によって求めたいのは，マイク j で観測される，対応する話者 j の音声であるから，これを

$$Y_j(f,t) = G_{jj}(f,t) S_j(f,t) \tag{4}$$

とおき，伝達関数を

$$H_{ij}(f,t) = \frac{G_{ij}(f,t)}{G_{jj}(f,t)} \tag{5}$$

によって置き換えると，観測信号は

$$X_i(f,t) = Y_i(f,t) + \sum_{j \neq i} H_{ij}(f,t) Y_j(f,t) \tag{6}$$

と表される．

観測信号のパワースペクトルは，

$$\begin{aligned} &|X_i(f,t)|^2 \\ &= \left| Y_i(f,t) + \sum_{j \neq i} H_{ij}(f,t) Y_j(f,t) \right|^2 \\ &= |Y_i(f,t)|^2 + \sum_{j \neq i} |H_{ij}(f,t) Y_j(f,t)|^2 \\ &\quad + \sum_{k=1}^{N} \sum_{j \neq k} |H_{ik}(f,t) Y_k(f,t) H_{ij}(f,t) Y_j(f,t)| \cos\theta_{kj,i} \end{aligned} \tag{7}$$

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．

*Speech separation based on spectral subtraction for meeting speech recognition. by Yu Nasu, Koichi Shinoda, and Sadaoki Furui (Tokyo Institute of Technology)

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

$$S_j(f,t)S_k(f,t) \approx 0 \tag{8}$$

が成り立つとすると，式 (7) の第 3 項は十分小さく無視できる．従って，目的信号は式 (2) と同様にして

$$|\hat{Y}_i(f,t)|^2 = |X_i(f,t)|^2 - \alpha \sum_{j \neq i} |\hat{H}_{ij}(f,t)|^2 |\hat{Y}_j(f,t)|^2 \tag{9}$$

と推定できる．

実際には $|\hat{H}_{ij}(f,t)|^2$ と $|\hat{Y}_j(f,t)|^2$ が未知であるため，これらを推定する必要がある．

3.2 伝達関数の推定

まず，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．会話音声においては，このようなフレームの存在は妥当な仮定である．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，

$$|\hat{Y}_i(f,t)|^2 = \max\left(|X_i(f,t)|^2 - \sum_{j \neq i} |X_j(f,t)|^2, 0 \right) \tag{10}$$

として，適当な閾値 $T_{j1}(t), T_{k2}(t)$ を用いて

$$\frac{1}{|F|} \sum_{f \in F} |\hat{Y}_j(f,t)|^2 > T_{j1}(t) \tag{11}$$

$$\frac{1}{|F|} \sum_{f \in F} |\hat{Y}_k(f,t)|^2 < T_{k2}(t), \quad \forall k \neq j \tag{12}$$

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．

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 (6) より

$$X_i(f,t) = \begin{cases} Y_j(f,t) & \text{if } i = j \\ H_{ij}(f,t)Y_j(f,t) & \text{otherwise} \end{cases} \tag{13}$$

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，忘却係数を $\rho_h \in [0,1]$ として，話者 j のみが発話しているフレームで

$$|\hat{H}_{ij}(f,t)|^2 = \rho_h |\hat{H}_{ij}(f,t-1)|^2 + (1-\rho_h)\frac{|X_i(f,t)|^2}{|X_j(f,t)|^2} \tag{14}$$

のように更新する．

3.3 分離信号の推定

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

$$|\hat{Y}_i^{(n)}(f,t)|^2 = |X_i(f,t)|^2 - \alpha_n \sum_{j \neq i} |\hat{H}_{ij}(f,t)|^2 |\hat{Y}_i^{(n-1)}(f,t)|^2 \tag{15}$$

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．この操作による推定値は連立方程式として直接的に解くよりも安定しており，歪みの小さい推定が可能である．

4 評価実験

複数の話者によって断続的に発話される音声の認識をタスクとして提案法の評価を行った．評価データとして，日本語話し言葉コーパス（CSJ）[7] の音声を使用した計算機シミュレーション，および実際に収録した会議データを用いた．

提案手法において，1 人だけが発話しているフレームの推定では，閾値を

$$T_{j1}(t) = \frac{2}{|F|} \sum_{f \in F} |\hat{N}_j(f,t)|^2 \tag{16}$$

$$T_{k2}(t) = \frac{1}{|F|} \sum_{f \in F} |\hat{N}_k(f,t)|^2 \tag{17}$$

とした．伝達関数の推定は 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 1,024 点で行った STFT の周波数を線形軸上で 64 分割して計算し，更新は $\rho_h = 0.98$，$|\hat{H}_{ij}(f,0)|^2 = 0.10$ として行った．

また予備実験の結果，反復の回数は 2 回で十分な効果が得られ，3 回以上に増やしても大きな変化はなかったため，実験では 2 回の反復を行い，減算係数を $\alpha_1 = 1.0, \alpha_2 = 4.0$ とした．

4.1 計算機シミュレーションによる評価

4.1.1 実験条件

CSJ のテストセットから男性話者 4 人の模擬講演音声を使用し，会議を模して各話者が断続的に発話する 7 分間のデータを作成した．4 人の話者の音声を 120 秒間ずつ用い，1.2–7.6 秒（平均 2.7 秒）の発話単位に区切って時間内にランダムな間隔で並べた．延べ形態素数は 1,593 で，各話者の音声区間長

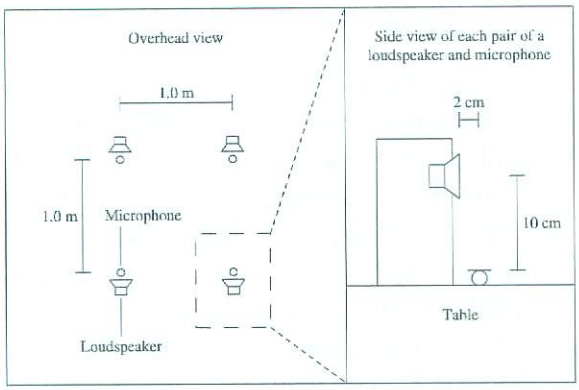

Fig. 1 Arrangement of loudspeakers and microphones

Table 1 Recognition results of simulated speech

	Baseline	Binary masking	Proposed	Ideal
Corr. [%]	75.3	78.3	82.6	84.4
Acc. [%]	66.5	71.5	77.7	79.0

は延べ 480 秒間，そのうち他者の音声区間が重なっている割合は 58 % である．標本化周波数は 16 kHz とした．

このデータに対し，インパルス応答の畳み込みと背景雑音の重畳を行って評価用データとした．インパルス応答は残響時間（RT_{60}）が 0.4 秒の会議室において，卓上に 4 本の単一指向性ピンマイクと 4 台のスピーカーを Fig. 1 のように配置して測定したものを用いた．背景雑音は無人の会議室において同じマイク配置で録音したものを用い，正解音声区間における平均 SNR が 25 dB となるように重畳した．

評価尺度には音声認識結果の単語正解率（Corr.）および単語正解精度（Acc.）を用いた．音響特徴量には MFCC 12 次元とその Δ, ΔΔ と，パワーの Δ, ΔΔ の計 38 次元を用い，ケプストラム平均除去を行った．音響モデルは 32 混合 3,000 状態 triphone HMM とし，CSJ の男性話者による模擬講演データから学習したものを使用した．評価データに用いた話者は学習データに含まず，音源分離により生じる歪みに対する適応も行っていない．実験には HTK-3.4.1 [8] を用いた．発話区間は既知として正解音声区間により分割し，4 人の話者の全発話について評価した．

4.1.2 実験結果

評価対象には，提案手法による処理音声（Proposed）のほか，比較対象として各マイクの観測音声をそのまま用いたもの（Baseline），および従来のスパース性に基づくバイナリマスクによる音源分離音声（Binary masking）を用いて実験を行った．従来法は各時間周波数における信号成分を，観測されたパワーが最大となるマイクに割り当てるアルゴリズム [4] により実装した．提案手法および従来手法の STFT はフレーム長 1,024 点，フレームシフト 512 点，Hamming 窓により行った．

評価結果を Table 1 に示す．ここでは，他者の音声と背景雑音の重畳を行わずに作成した各話者単独の音声（Ideal）の認識結果も併記した．正しい発話区間による切り出しが行われていても，観測音声をそのまま認識すると他者の音声の混入により単語正解精度が低下している．従来手法では 5.0 ポイントの改善がみられ，一定の効果があった．これに対し，提案手法では 11.2 ポイントの改善となり，単独発話音声に迫る高い精度が得られた．バイナリマスクを用いる従来手法に対し，提案手法では分離音声の歪みが改善されたためと考えられる．

4.2 会議データによる評価

4.2.1 評価データ

前節と同じ会議室において，男性話者 4 人による 20 分間の模擬会議を収録した．各話者の座席は Fig. 2 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．使用したマイクは前節と同じ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．音声区間および音声認識の正解ラベルには人手によって作成されたものを用いた．延べ形態素数は 5,154 で，各話者の音声区間長は延べ 1,496 秒間，そのうち他者の音声区間が重なっている割合は 47 % であった．なお，ここでは話者の笑い声や咳も音声区間に含めている．音声認識の条件は前節と同じとした．

4.2.2 実験結果

評価結果を Table 2 に示す．単語正解精度は最大でも 4 割に満たない低い値であった．収録した模擬会議では発話が極めて自発的であり，また発話の割り込みによる不完全な文や発音の怠けが頻繁に生じることから，仮に分離が完全に行われていても音声認識として困難なタスクであるためと考えられる．提案手法では観測音声より 6.3 ポイント高い単語正解精度が得られ，実環境においても従来手法を上回る効果が確認された．

4.3 立ち会議データによる評価

4.3.1 評価データ

前節と同じ会議室において，男性話者 4 人による 20 分間の模擬会議を収録した．各話者は前節と同じ

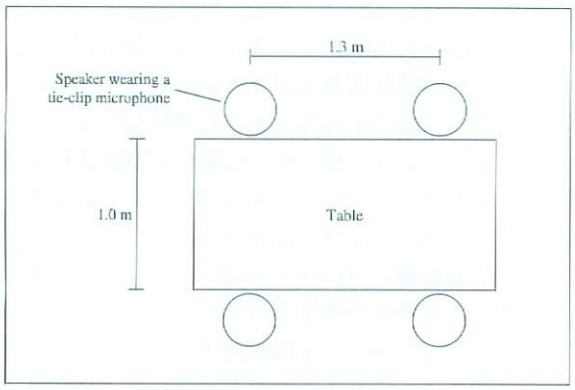

Fig. 2 Position of each speaker in sit-down meeting

Table 2 Recognition results of sit-down meeting speech

	Baseline	Binary masking	Proposed
Corr. [%]	40.2	41.2	43.9
Acc. [%]	30.6	32.1	36.9

Table 3 Recognition results of stand-up meeting speech

	Baseline	Binary masking	Proposed
Corr. [%]	45.1	44.3	47.9
Acc. [%]	37.5	35.2	40.6

単一指向性ピンマイクを胸元に装着し，ホワイトボードの前に立った状態でそれぞれ 1.5 m ほど離れて会議を行った．話者のうち 1 人は，会話をしながらホワイトボードにメモを書く役割を担当した．座席がないため，前節の会議と比較して話者の位置の変化は大きい．音声区間および音声認識の正解ラベルには人手によって作成されたものを用いた．延べ形態素数は 3,980 で，各話者の音声区間長は延べ 1,189 秒間，そのうち他者の音声区間が重なっている割合は 28 % であった．話者の笑い声や咳も音声区間に含めている．音声認識の条件は前節と同じとした．

4.3.2 実験結果

評価結果を Table 3 に示す．提案手法では観測音声より 3.1 ポイント高い単語正解精度となり，話者の位置変化が大きい会議においても有効であることが示された．

5 まとめ

会議音声認識を目的とするスペクトル減算に基づく音源分離手法を提案した．計算機シミュレーションおよび実際の会議データに対し，音声認識の実験により提案手法の有効性を確認した．シミュレーションでは高い効果があったが，実際の会議においては，従来法より大きい改善がみられたものの，単語正解精度は 40 % 程度にとどまった．自然な会話音声には他者の音声の混入以外にも音声認識が困難となる課題が多く，多方面からのアプローチが必要である．

本稿では通常の音響モデルを用いて音声認識を行ったが，歪みが含まれる分離音声を用いた学習や適応によって認識精度の向上が期待できるため，今後の課題としたい．

謝辞 本研究は科学研究費補助金基盤研究（B）20300063 の援助を受けた．

参考文献

[1] Shriberg *et al.*, Proc. Eurospeech, pp.1359–1362, 2001.
[2] Bell and Sejnowski, Neural Computation, vol.7, no.6, pp.1129–1159, 1995.
[3] Smaragdis, Neurocomputing, vol.22, pp.21–34, 1998.
[4] Aoki *et al.*, Acoust. Sci. Tech., vol.22, no.2, pp.149–157, 2001.
[5] Rickard *et al.*, Proc. ICA, pp.651–656, 2001.
[6] Boll, IEEE Trans. ASSP, vol.27, no.2, pp.113–120, 1979.
[7] Maekawa *et al.*, Proc. LREC, vol.2, pp.947–952, 2000.
[8] Hidden Markov Model Tool Kit (HTK), `http://htk.eng.cam.ac.uk/`.
[9] 那須 他，信学技報，vol. 110，no. 56，pp. 7–12，2010.