ロボット聴覚のためのソフトマスク生成法による周辺話者音声認識率の改善

○高橋 徹 (京都大学) 中臺 一博 (HRI-JP) 駒谷 和範 (京都大学)
尾形 哲也 (京都大学) 奥乃 博 (京都大学)

Improving Speech Recognition of Periphery Talkers by Generating Soft Masks for Robot Audition

*Toru TAKAHASHI (Kyoto Univ.), Kazuhiro NAKADAI (HRI-JP), Kazunori KOMATANI (Kyoto Univ.), Tetsuya OGATA (Kyoto Univ.), Hiroshi G. OKUNO (Kyoto Univ.)

Abstract— This paper addresses automatic soft missing feature mask generation based on leak energy estimation for a simultaneous speech recognition system. To realize a robot audition for automatic speech recognition equivalent to audition of people who recognize, it is helpful to use reliability of spectral representation. We should take into account reliability of spectral representation for the probability calculation in recognition process. The reliability generating method is designed. As itt is represented continuous number which ranges from 0.0 to 1.0, the reliability is named soft mask.
Experiment for recognising words which are simultaneously uttered by three speakers was conducted. The average recognition ratio based on soft missing feature mask was improved about 5 % for all direction from a conventional system based on hard missing feature mask. Word recognition ratio was also improved from 93 % to 97 % for front speech when speakers located apart from 90 degree.

Key Words: Robot Audition, Soft mask, Missing feature theory, Automatic speech recognition

1. はじめに

人間は, 様々な状況で音声を聴くことができる. 周囲に雑音があっても, 目的音声に注意を向け, 聞き取り, 理解する能力をもっている. 我々は, この能力を日常的に使っており, 例えばエアコン・コピー機の音・他の人の声など様々な音が存在しているオフィス内で会話できる. 他の話し声に邪魔されることなく, 相手の声に注意を向け、聞き取り、理解している [1]. 我々の目的は, この能力をロボット聴覚として実現することである. この目的を重視する理由は, ロボットが実環境で活動する時, ロボットの持ち主の声を聞き分けて, 適切なインタラクションを行うには, 音声を聞き分ける能力が必須だからである.

著者らは, 複数音源の音声を同時認識させるためにミッシングフィーチャ理論に基く音声認識システムを開発している [2]. ミッシングフィーチャ理論に基く音声認識は, 歪んだ音響特徴量をマスクし, 歪のない特徴量だけで認識することにより, 歪に強い音声認識を行える. 複数話者の音声を同時認識するには, 各音源からの音の空間的スパースネスを仮定する. 音響特徴の周波数的・時間的スパースネスの仮定は不要である. 課題は, この仮定の下, 各音源からの混合音を分離することである. 空間的スパースネスの仮定から, 必然的に音源間の物理的距離が認識率を支配する. 本報告では, ロボットを中心に 2m の円周上に 3 つの音源を様々な角度に配置して, 認識実験を行う. 音源間の角度が狭い時により困難な認識となる. 混合音の分離には, マイクロフォンアレーと Geometric Source Separation を用いる. 分離音にはチャネル間の漏洩による歪があるため, ミッシングフィーチャ理論に基く音声認識と組み合わせるのが適当である. 我々のシステムには, 0, 1 のハードマスク処理 [5] が実装されている. 音声認識のためのマスク生成の先行研究 [3][4] では, ロボットに実装されていない. 我々は, マスク生成部分を改良し, 0〜1 の連続値を用いるソフトマスク処理を開発したので報告する. ソフトマスクは, ハードマスクに比べ, 歪んだ音響特徴量から認識の手がかりをより多く抽出できると期待できる. 音源分離部分で推定した漏洩エネルギと背景雑音エネルギから信頼度を計算し, ソフトマスクを生成する方法を開発した.

2. システム概要

複数話者による同時発話音声認識システム（図 1）は, コンピュータによる音環境理解 (Coputational Auditory Sean Analysis) に基き, 次の 3 つのコンポーネントから構成される.

1. 音源分離 (Sound Source Separation), ただし音源定位を含む.
2. ミッシングフィーチャ理論に基く自動音声認識 (Automatic Speech Recognition), 音響特徴量抽出を含む.
3. 自動ミッシングフィーチャマスク (Missing Feature Mask) 生成.

システムは 8ch マイクロフォンアレーを搭載したロボット (図 3) に実装されている. 図 1 では, 8ch マイクロフォンアレーを経て, 複数音源からの混合音が音源分離モジュールに送られる様子を表している. 音源分離モジュールでは, 音源分離の他に音源数と音源分離による分離歪を推定する. 推定された音源数の分離音を音声認識モジュールに, 推定された分離歪を MFM 生成モジュールに送る. 自動音声認識モジュールでは, 音源分離モジュールで分離した音声の音響特徴量を求め,

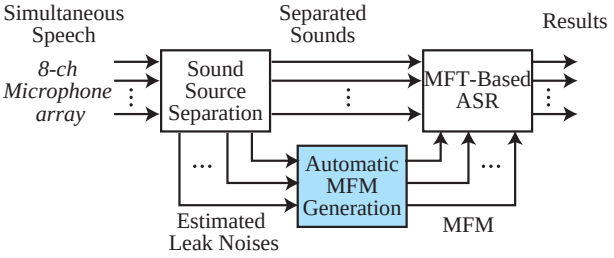

Fig.1 MFM-ASR 概要.

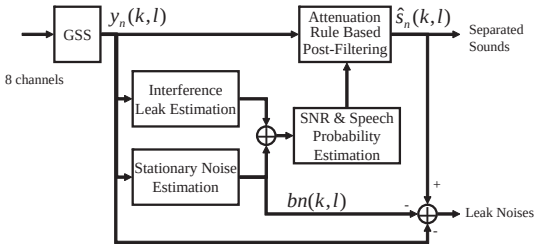

Fig.2 Geometric source separation with multi-channel post-filter.

MFM 生成モジュールで求めたマスクを使って音声認識を行う. 自動 MFM 生成モジュールは, 他の 2 つのモジュールを繋ぐ役割を担っている. このモジュールが無い場合は, 音響特徴量をマスクしないことと等価である. 分離音声から求めた音響特徴量を直接音声認識することに相当する.

2·1 音源分離

音源分離モジュールは, Geometric Source Separation (GSS)[6, 7, 2] とマルチチャネルポストフィルタ処理によって実装されている. 図 2 に概略を示す.

Parra [11] らの GSS アプローチをより高速に適応できるように改良している [2]. GSS によって得られる分離音は, 複数音源に対するビームフォマ・ポストフィルタに基くマルチチャネルポストフィルタ処理を経て最終的な分離音を求められる [2].

最初の分離音を強調するためのポストフィルタの係数は, 背景雑音と音源間の干渉を適応スペクトル推定値に基いて求められる.

このアプローチの基本的な特徴は, ノイズを定常要素とチャネル間の干渉による瞬時要素に分解することである.

GSS を用いたこの手法は, 周波数領域で処理される. 未知であるが実際の音源信号を $s_m(k,l)$ とする. m は音源番号, k は離散周波数のインデクス, l は時刻を表す.

音源 $s_m(k,l)$ に対応するベクトル は, $\boldsymbol{s}(k,l)$. 行列 $\boldsymbol{A}(k)$ は, 音源とマイク間の伝達関数を表す.

$$\boldsymbol{x}(k,l) = \boldsymbol{A}(k)\boldsymbol{s}(k,l) + \boldsymbol{n}(k,l), \tag{1}$$

ただし, $\boldsymbol{n}(k,l)$ は, 非コヒーレントな背景雑音である. 行列 $\boldsymbol{A}(k)$ は, 音源定位アルゴリズムで推定可能である. すべての伝達関数のゲインは, 一定であると仮定すると, 行列の要素は,

$$a_{ij}(k) = \exp\{-j2\pi k\delta_{ij}\} \tag{2}$$

と表せ, 分離結果は,

$$\boldsymbol{y}(k,l) = \boldsymbol{W}(k,l)\boldsymbol{x}(k,l), \tag{3}$$

と表せる. ただし, $\boldsymbol{W}(k,l)$ は分離行列を表す. Valin らの GSS 音源定位アルゴリズム [2] によって推定できる.

GSS 音源定位の出力は, Ephraim [12] らの最適推定値に基く周波数領域ポストフィルタで強調される. マルチチャネルポストフィルタの入力は, GSS の出力である $(\boldsymbol{y}(k,l) = (y_1(k,l), \cdots, y_M(k,l)))$. マルチチャネルポストフィルタの出力 $\hat{\boldsymbol{s}}(k,l)$ は,

$$\hat{\boldsymbol{s}}(k,l) = G(k,l)\boldsymbol{y}(k,l), \tag{4}$$

と表される. ただし, $G(k,l)$ は, ゲインである. $G(k,l)$ の推定値は, スペクトル振幅の最小二乗誤差基準で求める. $G(k,l)$ を求めるために, ノイズの分散が推定される.

ノイズの分散推定値 $\lambda_m(k,l)$ は,

$$\lambda_m(k,l) = \lambda_m^{stat.}(k,l) + \lambda_m^{leak}(k,l), \tag{5}$$

と表される. ただし, $\lambda_m^{stat.}(k,l)$ と $\lambda_m^{leak}(k,l)$ は, 音源 m のノイズの定常要素の推定値と音源の干渉の推定値である. l, k は, フレーム (時刻) と離散周波数のインデクスを表す.

定常雑音の推定値 $\lambda_m^{stat.}(k,l)$ は, Minima Controlled Recursive Average (MCRA) によって求める. λ_m^{leak} は, 他の音源からの干渉が, 要素 η によって減少 (典型的には -10 dB 〜 -5 dB) する仮定のもとで, 推定される. 干渉の推定値は,

$$\lambda_m^{leak}(k,l) = \eta \sum_{i=0, i\neq m}^{M-1} Z_i(k,l), \tag{6}$$

と表される. ただし, $Z_i(k,l)$ は, 音源 m の平滑化スペクトルで, スペクトル $Y_m(k,l)$ を用いて再帰的に定義される [2].

$$Z_m(k,l) = \alpha_s Z_m(k,l-1) + (1-\alpha_s)Y_m(k,l). \tag{7}$$

ただし, $\alpha_s - 0.7$ である.

2·2 ミッシングフィーチャ理論に基く自動音声認識

音声認識に用いる音響特徴量は, 耐雑音性が高く, 音源分離後の歪の影響が一部のパラメタに限定されるのが望ましい. 我々は, 山本らによって開発された Mel Scale Logarithmic Spectrum (MSLS) [6] を用いる. MSLS ベクトル表現を用いることで, 特定の周波数帯域の歪は一部の次元に表れ, マスクし易くなる. MSLS ベクトルは 48 次元ベクトルを用い, 低次の 24 次元をスペクトル特徴量に, 高次の 24 次元をスペクトルの時間差分特徴量に用いる.

ミッシングフィーチャ理論に基く自動音声認識 (Missing Feature Theory based Automatic Speech Recognition) は, 音響特徴量系列とマスク系列から 音素列を出力する. MFT-ASR モジュールは, 一般に広く用いられている HMM に基いている. 従来の HMM に基く ASR システムでは, 最尤の音素系列は, 状態遷移確率と出力確率から推定される. MFT-ASR では, 出力確率の推定プロセスが改良されている.

$x(i)$ を MSLS ベクトルの i 番目の要素とし, $M(i)$ をミッシングフィーチャマスク (MFM) ベクトルの i 番目の要素とする. $M(i)$ は, $x(i)$ の信頼度を表す.

出力確率 $b_j(x)$ は,

$$b_j(x) = \sum_{l=1}^{L} P(l|S_j) \exp\Big\{\sum_{i=1}^{N} M(i) \log f(x(i)|l, S_j)\Big\}, \tag{8}$$

と定義する．但し, $P(\cdot)$ は, 確率密度を表す．N は, 音響特徴量の次数である．S_j は, j 番目の状態を表す．$f(x|S_j)$ は, 状態 j の L 混合の N 次ガウス分布である．

音響特徴量の信頼度の知識が得られなければ, MFM はすべて 1 となり, 音響特徴量はそのまま認識に用いられる．この時, 従来の認識と等価となる．

2·3 自動 MFM 生成

48 次元 ($2N = 48$) のスペクトル特徴ベクトルを用いる．MFM は, ベクトルの各要素に対応した信頼度から成る特徴ベクトルで 0 の時信頼できないことを, 1 の時信頼できることを表す．従来 MFM は, 0 または 1 のバイナリマスクが用いられてきた．本報告では, 0 から 1 の連続値にマスクを拡張する．ソフトマスクは, Barker ら [13] によって音声認識に用いられており, Barker らの方法に倣いシグモイド関数を用いる．更に, 静的スペクトル特徴量と動的スペクトル特徴量の重みを導入する．

マスクは, 音源分離モジュールの処理過程で得られるパラメタから生成できる．音源分離モジュール中の GSS 処理によって分離された音声 $y_m(k,l)$ から, $\hat{s}_m(k,l)$ と背景雑音の推定値 $bn(k,l)$ を出力する．それぞれの信号をメルフィルタバンクに通した時の出力エネルギを $Y_m(k,l)$, $\hat{S}_m(k,l)$, $BN(k,l)$ と表す．各メルフィルタバンクに対し, 特徴量は入力エネルギと出力エネルギの比がしきい値 T_{MFM} を越える時, 信頼できる．これは, より多くのノイズが存在する周波数帯では, ポストフィルタ利得が少いという仮定の上に成り立っている．続いて, 信頼度を定義し, 従来方であるハードマスクの定義を示した後, 提案手法のソフトマスクの定義を示す．

2·3.1 ハードマスク (従来のマスク生成法)

信頼度を $R(k,i)$ と表すと, 静的スペクトル特徴量のハードマスク $M_m^H(k,i)$, $(i = 1, \cdots, N)$ と動的スペクトル特徴量のハードマスク $M_m^H(k,i)$, $(i = N{+}1, \cdots, 2N)$ は,

$$M_m^H(k,i) = \begin{cases} 1, & R(k,i) > T_{MFM} \\ 0, & \text{otherwise} \end{cases}, \quad (9)$$

$$M_m^H(k,i) = \prod_{j=k-2, j\neq k}^{k+2} M_m(j, i-N), \quad (10)$$

$$R(k,i) = \frac{\hat{S}_m(k,i) + BN(k,i)}{Y_m(k,i)}, \quad (11)$$

と定義されている．動的スペクトル特徴量が隣接 2 フレーム以内の静的スペクトル特徴量差の平均で定義されているため, 動的スペクトル特徴量の信頼度は, 隣接 2 フレーム以内の静的スペクトル特徴量がすべて信頼できるときに信頼できると定義されている．

2·3.2 ソフトマスク (開発したマスク生成法)

静的スペクトル特徴量のソフトマスク $M_m^S(k,i)$, $(i = 1, \cdots, N)$ は,

$$M_m^S(k,i) = wQ(R(k,i)|a,b), \quad (12)$$

$$Q(x|a,b) = \begin{cases} \frac{1}{1+\exp(-a(x-b))}, & x > b \\ 0, & \text{otherwise} \end{cases}, \quad (13)$$

と定義する．ただし, w は, 結合重みを表す．$0.0 < w < 1.0$ の範囲で変えることで, 認識プロセスで動的スペクトル特徴量に比重を置いた音声認識を行える．$w > 1$ で静的スペクトル特徴量に比重が置かれる．

$Q(\cdot|a,b)$ は, 2 つのパラメタをもったシグモイド関数で, a, b は, それぞれ関数の傾斜と位置に対応している．$w = 1$ で ,a を十分大きくとると, ソフトマスクは, ハードマスクに漸近する．この時, b は, ハードマスクのしきい値 T_{MFM} とみなせる．

静的スペクトル特徴量と動的スペクトル特徴量の評価重み w を導入する理由は, 静的スペクトル特徴量に比べ, 動的スペクトル特徴量がより音源分離による歪の影響を大きく受けるため, 動的スペクトル特徴量部分の音響尤度が, マスクによる改善される幅が小さいことを補償するためである．

動的スペクトル特徴量は, いくつかの連続する静的スペクトル特徴量から求めるため, 動的スペクトルのソフトマスクは, 静的スペクトル特徴量の信頼度の積で定義されている．

$$M_m^S(k,i) = \prod_{i=k-2, i\neq k}^{k+2} Q(R(k,i|a,b)). \quad (14)$$

信頼性計算の範囲に, 信頼性の低い静的スペクトル特徴量があると, 信頼度の積で表されているため, 信頼性が低くなる傾向がある．一般に, 動的スペクトル特徴量の信頼性は, 静的スペクトル特徴量の信頼性に比べて低くなる．

従って, 例えソフトマスクによって, 動的スペクトル特徴量の音響尤度が改善されても一発話全体の音響尤度の寄与率は小さい．動的スペクトル特徴量に比重を置くことでマスク処理による小さな尤度変化を全体の音響尤度に占める割合を大きくできる．これにより, 認識結果の改善を促せる．

3. 実験

提案手法に基く自動 MFM 生成によって生成されたマスクを評価するために, 3 話者同時発話単語認識実験を行った．

実験には 左右対象に配置した合計 8 個のマイクを搭載した SIG2 を用いた．マイクはロボットの身体に密着しているため, マイクの周波数特性は, ロボット身体の影響を受ける．予め測定した身体の伝達関数によって影響を補正した．マイクの設置場所を図 3 に示す．

ラウドスピーカをロボット正面に 1 つ, 左右に 2 つ対象に配置した．左右の位置は, 10 度間隔で 90 まで条件を変て実験を行った．実験環境を図 4 に示す．残響時間約 0.35 秒 (RT20) の部屋でロボットから 200cm に配置した 3 つのラウドスピーカから異なる単語を同時に再生し, 評価実験を行った．評価に用いた単語は, ATR 音素 216 バランス単語中の 200 単語である．

MFT-ASR として, マルチバンド版 Julius [8, 9, 10] を使った．言語モデルは文法ベースで, 音響モデルは, トライフォンモデルを用いた．

3·1 実験結果と考察

3 話者同時発話における単語認識実験を行った．図 5 は, ハードマスクとソフトマスクを使った場合の単語正解率を表す．この結果は, 表 1 に示すマスク生成におけるパラメタ中で最高の正解率を表わすパラメタ値を用いている．横軸は, 3 話者 (実際にはラウドスピーカーを配置) の位置を表す．1 名は常に正面に立ち, 残る 2 名は左右に 30, 60, 90 度の位置に立つ場合をそれ

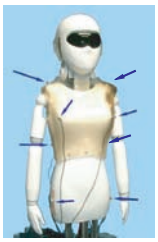

Fig.3 *Humanoid robot SIG2 and location of eight microphones.*

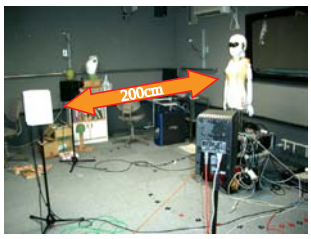

Fig.4 *Humanoid robot SIG2 and location of speakers.*

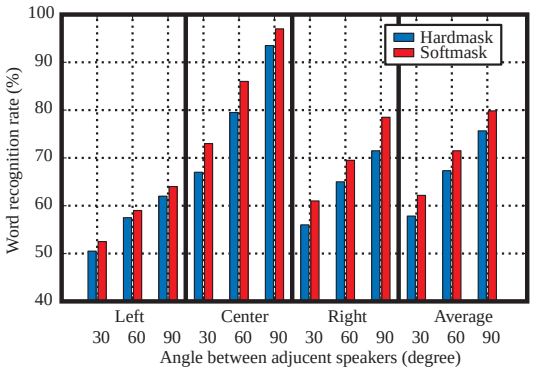

Fig.5 Word recognition ratio for each direction.

Table 1 マスク生成のためのパラメタ探索範囲

Parameter	ハードマスク	ソフトマスク
しきい値 T_{MFM}	0.05–0.35	–
傾き a	–	60–140
中心 b	–	0.05–0.35
重み w	–	0.4

ぞれ 30, 60, 90 と表す. 正面の話者を Center, 左右をの話者を Left, Right と表す. 例えば, 「30 Left」の項目は, ロボット正面から左側 30 度の位置に立つ話者の単語正解率を表している. Average の項目は, 正面と左右の立ち位置について平均した結果である.

単語正解率が, 話者間の角度によらず約 5 % 改善されている. これは, ハードマスクでは信頼できなものとして無視されていた音響的手がかりをソフトマスクを導入することによって, 利用できた結果である.

角度が狭くなると単語正解率が低下する傾向がある. これは, 空間的スパースネスを仮定しているので角度が狭くなることで音源分離の漏洩歪が大きくなり, 信頼できる音響特徴量が減り, 音響特徴を区別するのが困難になった結果であると考えらえる.

左の話者の単語正解率が, 他の方向の話者に比べて低く, 改善幅も小さいことの理由は, わかっていない. 他の話者セットで評価することによって話者特有の問題かどうか確かめる必要がある.

4. 結論

ロボット聴覚のためのソフトマスク生成法を開発し, 3 話者同時発話音声認識実験を行い効果を確認した. ハードマスクとソフトマスクを比較し, 平均単語正解率が 5% 改善し, 正面の発話者で 93 % から 97 % に改善した. 本実験による最大の正解率の向上は, 右側の話者で 8 % であり, 特に周辺話者の正解率の向上に効果がある点が重要である. なぜなら, SIG2 に配置されたマイクが正面に対して分離性能が最大になる位置に配置され, 横方向が最も分離が困難だからである. 一方, 正面話者に注目すると, 左右 30 度の位置にそれぞれ発話者がいるにもかかわらず, 70 % 以上の正解率を達成できた. 60 度の場合では, 約 85 % を達成している. すべての話者の正解率を改善することと同様に, 周辺に複数の話者がいるときの, 正面話者の正解率を改善することも重要である. 今後, この二つの課題を検討していく予定である.

5. 謝辞

本研究は, 京都大学グローバル COE プログラム, 科研費基盤 (S) の支援により行われた.

参考文献

[1] M. Kashino et al., "One, two, many – judging the number of concurrent talkers", *Journal of Acoustic Society of America*, vol. 99, no.4, 1966, pp.Pt.2, 2596.

[2] S. Yamamoto et al., "Genetic algorithm-based improvement of robot hearing capabilities in separating and recognizing simultaneous speech signals", *Proc. IEA/AIE* Vo. LNAI 4031, 2006, pp.207-217, Springer-Verlag.

[3] M. L. Seltzer et al., "A bayesian framework for spectrographic mask estimation for missing featuer speech recognition", *Speech Communication*, vol.43, 2004, pp.379–393.

[4] Raj Bhiksha et al., "Missing-Feature Approaches in Speech Recognition", *Sigral Processing Magazine*, vol.22, no. 5, 2005, pp.101–116.

[5] S. Yamamoto et al., "Improving Speech Recognition of Simultaneous Speech Signals by Parameter Optimization with Genetic Algorithm", *Proc. International Conf. on Robotics and Automation*, 2006.

[6] S. Yamamoto et al., "Enhanced robot speech recognition based on microphone array source separation and missing feature theory", *Proc. International Conf. on Robotics and Automation*, 2005, pp.1489–1449.

[7] S. Yamamoto et al., "Making a robot recognize three simultaneous sentences in real-time", *Proc. International Conf. on Intelligent Robots and Systems*, 2005, pp.897–902.

[8] Y. Nishimura et al., "Noise-robust speech recognition using multi-band spectral features", *Proc. 148th Acoustical Society of America Meetings*, No. 1aSC7, 2004.

[9] Multiband Julius, `http://www.furui.cs.titech.ac.jp/mband_julius/`

[10] T. Kawahara et al., "Free software toolkit for Japanese large vocabulary continuous speech recognition", *Proc. of International Conf. on Spoken Language Processing*, vol.4, 2000, pp.476–479.

[11] Lucas C. Parra et al, "Geometric Source Separation: Merging Convolutive Source Separation With Geometric Beamforming," *IEEE Trans. Speech and Audio Processing*, vol.10, no.6, pp.352–362, 2002.

[12] Y. Ephraim et al., "Speech enchancement using minimum mean-square error log-spectral amplitude estimator", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-33, no.2, 1985, pp.443–445.

[13] J. Barker et al, "Soft decision in missing data techniques for robust automatic speech recognition," *Proc. International Conf. on Spoken Language Processing*, 2000.