ロボットとの対話における３マイクを用いた発話者の位置推定法

玉川　聡　　山本　一公　　中川　聖一

豊橋技術科学大学 工学研究科 情報・知能工学専攻

Estimation of Speaker's Position using Three Microphones in Spoken Dialog System with the Robot

Akira Tamagawa　　Kazumasa Yamamoto　　Seiichi Nakagawa

Computer Science and Engineering, Toyohashi University of Technology

1.　はじめに

本稿では，ロボットが搭載する 3 マイクを用いた方向・距離推定法を提案する．

方向推定は，各マイクの音声の到達時間差（TDOA）に基づく発話者の空間位置推定式より行う．距離推定は，上記方向推定結果に，予め収録データにより作成した方向別の TDOA を入力特徴量とする SVM 識別器を用いて，50cm と 100cm の距離判別を行う．

実験の結果，方向推定は最大許容誤差±15° で 82%，許容誤差±30° で 94%の精度であった．この推定結果を用いた距離判別は全方向で 68%，前方向（-60° から +60° ）のみだと 83%の精度が得られた．

本稿の構成は以下のようになる．2 節で今回実装した対話システムの概要を述べ，3 節で方向推定，4 節で距離推定の考え方とその方法を述べる．5 節で実験内容とその結果を示し，6 章はまとめと今後の課題について述べる．

2.　ロボットとの音声対話システムの概要

人間と対話を行うコンピュータやロボットの実現は，人工知能研究において大きな課題の一つである．国内でも多くの音声対話システムが存在し，奈良先端科学技術大学院大学が開発したロボットによる駅周辺の音声案内システム「キタちゃんロボット」[1]や，名古屋工業大学の開発した大型ディスプレイによる情報案内システム「メイちゃん」[2]などがある．企業開発としては，日本電気株式会社（NEC）が開発した音声認識・音声合成によるコミュニケーションロボット「PaPeRo」[3]や，株式会社東芝が開発した音声指示を習得して家電・AV 機器を操作するインタフェースロボット「ApriPoco™」[4]などがある．

本研究ではこのような背景をふまえ，音声対話システムの構築を行う．本研究の目的はユーザが話しやすい・使いやすい・親しみやすいシステムを開発することであり，ユーザが対話を楽しみながらタスクを達成することを目的とする．このような観点から，多機能ペットロボット Phyno を話し相手とした対話システムを構築する．

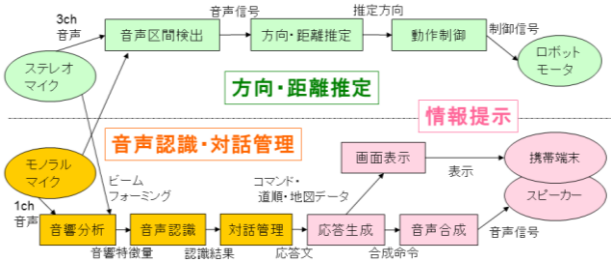

図 1. システムの流れ

本システム[5]のタスクは建物内の部屋へのルート案内とし，設置場所を本学研究棟 1 階のエレベータ前とした．図 1 に示すように，システムの構成は方向・距離推定機能，音声認識・対話管理機能，情報提示機能の 3 つに分けられる．

方向推定機能は，さらに細かく音声区間検出部，方向・距離推定部，動作制御部に分かれる．方向推定部では，各マイク間の音声の到達時間差（TDOA: Time Difference of Arrival）より，ユーザの音声の到来方向を推定し，動作制御部でロボットが音声の到来方向を向く．距離推定部では，ユーザがロボットに近いか（50cm），遠いか（100cm）を判別し，遠い場合はもっと近づいてもらうか，大きな声で発声するように呼びかける．

音声認識･対話管理機能は，さらに細かく音響分析部，音声認識部，対話管理部に分けられる．音響分析部では遠隔発話のためのビームフォーミング技術の導入を考えている．音声認識部は本研究室で開発された文脈自由文法（CFG）駆動型連続音声認識システム SPOJUS[6][7]を用いている．対話管理部は音声認識結果をもとに対話内容の遷移を管理する．

情報提示機能は，応答生成部，音声合成部，画面表示部からなる．応答生成部は各発話毎に適切な応答文を用意する．音声合成ではオープンソースで無償利用できる日本語の音声合成エンジン GalateaTalk[8]を使用する．GalateaTalk のコア部分は gtalk と呼ばれるが，本システムでは高速性，安定性に重きをおいて開発されている gtalk の派生版の jagtalk[9]を用いている．画面表示部では音声案内のみであることによる道順の不理解を解消するため，携帯端末に案内内容を表示させるための制御をする．

本システムの使用場面を図 2 に示す．

図 2. システムの使用場面

3.　方向推定

本システムで用いるロボット Phyno には，両耳の無指向性マイク（L,R）と口元の指向性マイク（C）が搭載されている（図 3 参照）．この内，両耳のマイクは本システムで方向推定をするために追加で搭載したものである．

方向・距離推定ではこの 3 マイクを全て用いる．方向推定では，人間と対話する上で支障が出ない程度だと想定される 30° 刻みで推定を行うものとする（図 4 参照）．

図 3. ペットロボット Phyno の外観と機能

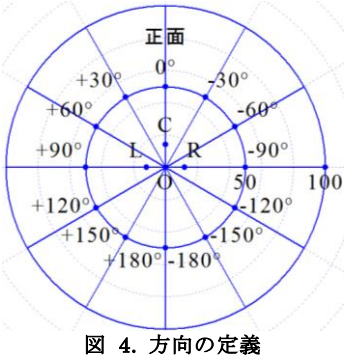

図 4. 方向の定義

3. 1　音源定位の原理

測定誤差がない場合，TDOA に基づく三角測量法により，2 マイクのとき方向推定可能であり（ただし，前後は不可），3 マイクのとき高さが既知なら音源位置推定可能，4 マイクのとき音源位置が推定可能である[10]．

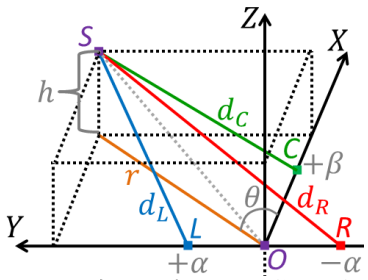

図 5. 発話者とロボットとの位置関係

図 5 のように，ロボットの中心Oと発話者の音源Sの水平面上の距離をr，音源 S とマイクiの高さの差をh，方向 θ から発話したとし，音源座標 S を(x_S, y_S, z_S)，マイクi の座標を(x_i, y_i, z_i)とすると，マイクi と音源Sの距離d_iと，各マイクi, jと音源S 間の距離差δ_{ij} は以下のようになる．

$$\mathrm{d_i} = \sqrt{(x_S - x_i)^2 + (y_S - y_i)^2 + (z_S - z_i)^2} \quad (1)$$

$$\delta_{ij} = \mathrm{d_i} - \mathrm{d_j} \quad (2)$$

本実験では$z_S - z_i$ を既知としたので，δ_{ij} を各マイク毎に求めれば，理論的に音源定位が可能である．今回の実験において，左マイクL の座標は$(x_L, y_L, z_L) = (0, \alpha, 0)$，右マイク$R$ の座標は$(x_R, y_R, z_R) = (0, -\alpha, 0)$，中央マイク$C$ の座標は$(x_C, y_C, z_C) = (\beta, 0, 0)$，$\alpha = 7.6$ [cm]，$\beta = 8.4$ [cm]，音源の高さ$h = 30$ [cm]（ただし実際はユーザの身長差によって可変であり 30～45cm 程度の範囲）である．

また，サンプリング周波数$f = 48000$ [Hz]，音速 34 [cm/msec]とすると，距離分解能は最大 0.71 [cm/sample]，角度分解能は 2.7° である．さらに，音速差や身長差，測定誤差によって，正確な位置推定は困難であり，実際に実験でも位置推定までは不可能であった．しかし，実験での間違った推定位置から音声到来方向を算出したところ，全方向で 8 割以上の精度で方向推定が可能であることがわかった．そこで，方向推定の第 1 段階として 2 マイクで困難な前後判別を 3 マイクの TDOA より求め，第 2 段階で 2 マイク（L,R）の TDOA より正確な方向推定を行うという手法をとった．

3. 2　GCC-PHAT 法

本システムでは，各マイク間の TDOA を GCC-PHAT(Generalized Cross-Correlation methods with Phase Transform)法を用いて求めるために，[11][12]で用いられているプログラムを使用する．

GCC-PHAT 法とは，周波数領域の計算によって 2 つの音声波形の相互相関関数を求める方法である．2 つの音声の相互相関関数は時間領域では畳み込みによって求まる．しかし，畳み込みによる演算は計算量が多いため，周波数領域に変換することで畳み込みの演算が掛け算で済み，計算量を大幅に減らすことができる．

一方のマイクiの音声波形を$x_i(t)$，他方のマイクjの音声波形を$x_j(t)$とするとき，それぞれ短時間スペクトル$X_i(f)$，$X_j(f)$に変換する．そしてこれらを掛け合わせたのち，時間領域に戻すことで相互相関関数$R(\tau_{ij})$を求める．$R(\tau_{ij})$は以下の式で表すことができる．ここで，$X^*_j(f)$は$X_j(f)$の複素共役である．

$$R(\tau_{ij}) = \sum_f \frac{X_i(f)X^*_j(f)}{|X_i(f)X^*_j(f)|} e^{-j2\pi\tau} \quad (3)$$

そして，相互相関関数$R(\tau_{ij})$の値が最大となるときのτ_{ij}がマイクi，jの音声波形の時間差（TDOA）$\hat{\tau}_{ij}$となる．これは(4)式より求まる．

$$\hat{\tau}_{ij} = arg \max_{\tau_{ij}} R(\tau_{ij}) \quad (4)$$

また，[11][12]で使用されているプログラムでは，使用する周波数帯域を自動的に更新するようにマスクがかけられる．プログラムでは，TDOA $\hat{\tau}_{ij}$の代わりにサンプリングポイント差$p_{ij} = f\,\hat{\tau}_{ij}/1000$ が出力される．

4.　距離推定

距離推定は原理的には各マイクの TDOA より解析的に求められるが，高さhの誤差や TDOA の測定誤差，音速差などのため，正確には求められなかった．そこで，事前に収録したデータより方向別に SVM 識別器を構成し，50cm と 100cm の 2 分類問題とした．

4. 1　距離推定の原理

同じ方向において，距離r_1，r_2の距離識別は，ポイント差の差$p_{ij,r2-r1} = p_{ij,r2} - p_{ij,r1}$が 1 ポイント以上あれば可能である．3 節の条件のとき，ポイント差p_{ij}の差$p_{ij,100-50}$は表 1 のようになり，理論的に距離判別は可能であることがわかる．表以外の方向についても同様に距離判別可能である．

表 1. ポイント差の差 $p_{ij,100-50}$の理論値

方向＼i,j	L,R	L,C	R,C
0°	+0.00	+1.10	+1.10
-30°	+1.12	+1.59	+0.47
-60°	+1.91	+1.71	-0.20
-90°	+2.19	+1.31	-0.88

4. 2　距離と各特徴量の関係性

実験時に収録した実環境データより，方向別に抽出した各特徴量と話者との距離の相関を調べた結果を図 6 に示す．図 6 では各方向について，距離と各特徴量の相関係数を求め，その絶対値の平均を求めた（距離との相関が大きいほど値が大きくなる）．抽出した特徴量のうち，代表的な特徴量や距離と相関の高かった特徴量を以下に示す．

- P-LR：マイクL,R間のポイント差p_{LR}
- P-LC：マイクL,C間のポイント差p_{LC}
- P-RC：マイクR,C間のポイント差p_{RC}
- Pow-L：マイクLの音声波形のパワー
- Pow-lowL：マイクLの低域通過フィルタを通した波形のパワー
- Pow-hiL：マイクLの高域通過フィルタを通した波形のパワー
- Pow-L/R：マイクL,Rのパワー比
- Pow-(l-h L/R)：$\frac{\text{マイク}L\text{の高域と低域のパワー比}}{\text{マイク}R\text{の高域と低域のパワー比}}$
- Cor-LC：GCC-PHAT プログラム内で求まるマイクL,C間の音声波形の相関のピーク値
- Cor-hiLR：マイクL,Rに高域通過フィルタを通した波形における，GCC-PHAT プログラム内で求まるの相関のピーク値

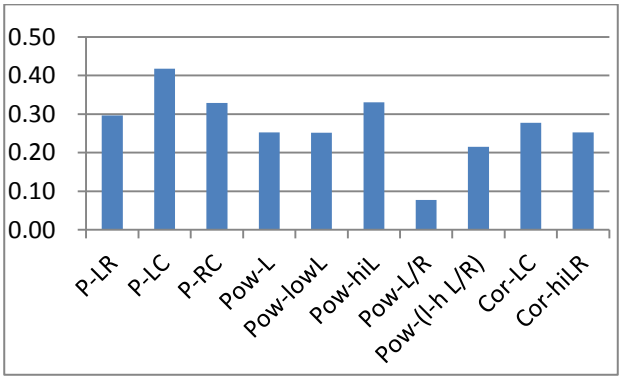

図 6. 実環境における各特徴量と距離の関係性

実験より，各マイクの音声の到達ポイント差 p_{ij}（図中 P-LR，P-LC，P-RC）のみを学習・テストデータに用いた場合が距離判別に一番有効であることがわかった．

5.　評価実験

5. 1　実験概要

実験に用いるデータは，防音室と実環境で収録した．実環境は，システムの利用を想定している本学研究棟 1 階のエレベータ前である．発話者は成人男性 6 人で，発話文は IPA100 文（毎日新聞）から選んだ 9 文を 3 セットに分け，各セットを 2 人ずつに 30° 刻みの方向別，2 距離別で発話してもらった．各環境における収録データは 468 発話文である．また，距離推定において，SVM 識別器の学習・テストには SVM^{light} [13]を用いた．

5. 2　方向推定実験結果

方向推定において，第 1 段階の 3 マイクの TDOA を用いた前後判別の正解率について図 7，図 8，表 3 に示す．

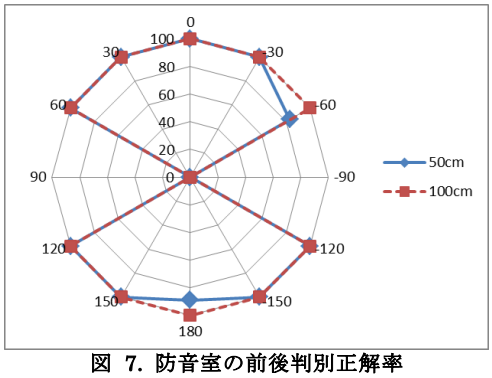

図 7. 防音室の前後判別正解率

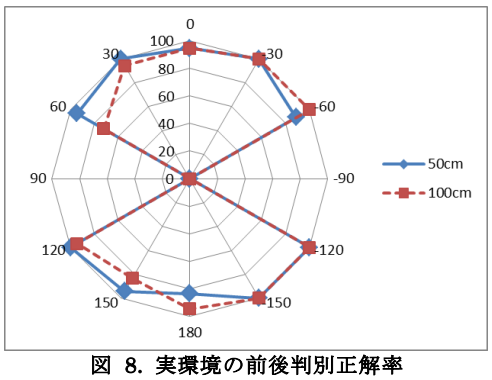

図 8. 実環境の前後判別正解率

結果より横方向から話しかけられるほど，前後判別が難しくなることがわかる．また，実環境において前後判別率が左右で偏っていることについては，実環境における遮蔽物の位置関係が関係していると考えられる．その要因の一つとして，ロボットの左側に壁があることでマイクが反射音を拾い，正解率に影響を及ぼしたことが考えられる．施設の位置関係を図 9 に示す．図におけるロボットと自動ドアの距離は約 435cm である．

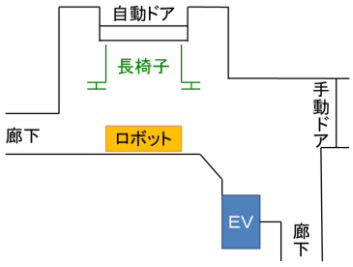

図 9. 実環境における施設の位置関係

次に方向推定の第 2 段階である 2 マイク（L,R）の TDOA による方向推定の結果を表 2 に示す．表中の「前方向」とは-60° から+60°，「後方向」とは-180° から -90° と+90° から+180°，「全方向」とは-180° から +180° の結果の平均値である．

表 2. 方向推定の実験結果（正解率）

方向	前方向		後方向		全方向	
許容誤差	±15°	±30°	±15°	±30°	±15°	±30°
防音室	96.7	100.0	93.4	99.3	94.7	99.6
実環境	82.8	93.9	81.3	93.4	81.8	93.6

結果より，方向推定は前後の判定を含め，防音室のと

き最大許容誤差±15° で 95%，許容誤差±30° で 99%以上の精度であり，実環境のとき許容誤差±15° で 82%，許容誤差±30° で 94%の精度であった．

5.3 距離推定実験結果

話者オープンの推定結果を出すため，収録データを学習用の 5 人分のデータ（390 発話文）とテスト用の 1 人分のデータ（78 発話文）に分け，反復的にテストを行った．本研究の方向の定義では，-180° と+180° は実環境でロボットのどちら側に壁があるかが違うため別々のデータとしているが，SVM 識別器の学習時は-180° と+180° を両方用いた方が性能が良くなった．

5.2 節における方向推定結果を用いて距離推定を行った結果を表 4 に示す．SVM 識別器において，方向の誤認識の対処・検証をするため，30° 方向の場合は学習・テストデータともに 30° と 60° 方向のデータを用いる，という様にしている．防音室の距離判別正解率は前方向（-60° から+60°)のみだと 91%，全方向で 79%であり，実環境では前方向（-60° から+60° ）のみだと 83%，全方向で 68%であった．また，表中の「学習防音室,テスト実環境」は，SVM 識別器の学習データには防音室のデータ，テストデータには実環境のデータを用いたものであり，つまり事前にその環境でデータを収録する必要がない環境非依存の性能となっている．結果より，環境非依存でも 70%程度の距離推定は可能だということがわかった．防音室の学習データを 15° 刻みにして識別器を学習させれば，よりよい性能が得られると考えられる．

6. むすび

本稿では，ロボットが搭載する 3 マイクを用いた方向・距離推定法を提案した．

方向推定は各マイクの音声到達時間差（TDOA）に基づく発話者の空間位置推定式より行い，距離推定は予め収録データで作成した方向別の SVM 識別器より 50cm と 100cm の距離判別を行った．実験の結果，方向推定は最大許容誤差±15° で 82%，許容誤差±30° で 94%の精度であり，この推定結果を用いた距離判別は全方向で 68%，前方向(-60° から+60°)のみだと 83%の精度が得られた．この精度なら想定している対話にも十分活用できると考えられる．

今後の課題として，30° 刻みでない発話データを収録し，30° 刻みで学習した SVM 識別器の識別率にどの程度影響が出るか調査したい．

参考文献

[1] 川波弘道,木田学,早川直樹,ツインツァレクトビアス,北村任宏,加藤智之,鹿野清宏: ”駅構内音声対話システム「キタちゃん」「キタちゃんロボットの開発 」” , IEICETechnicalReport, SP2006-14 (2006-06).

[2] 李晃伸，大浦圭一郎，徳田恵一: 魅力ある音声インタラクションシステムを構築するためのオープンソースツールキット MMDAgent，情報処理学会研究会，Vol.2011-SLP-89 No.27，(2011-12).

[3] コミュニケーションロボット PaPeRo, http://www.nec.co.jp/products/robot/

[4] インタフェースロボット ApriPoco™，東芝レビュー Vol.63 No.3，p.39，(2008).

[5] 玉川聡: ロボットとの対話による音声案内システムの開発，豊橋技術科学大学 卒業論文 (2010).

[6] 甲斐充彦,中川聖一:日本語連続音声認識システム SPOJUS-SYNO の改良と評価,電子情報通信学会技術報告 SP93-20 (1993).

[7] SPOJUS-SYNO, http://www.slp.ics.tut.ac.jp/SPOJUS/ .

[8] GalateaTalk プロジェクト, http://sourceforge.jp/projects/galateatalk/ .

[9] jagtalk [ja.nishimotz.com], http://ja.nishimotz.com/project:jagtalk .

[10] Longbiao Wang, Norihide Kitaoka, Seiichi Nakagawa: Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM, Speech Communication, Vol. 49, No.6, pp. 501-513 (Jun. 2007).

[11] Alberto Yoshihiro Nakano, Seiichi Nakagawa, Kazumasa Yamamoto: Automatic estimation of position and orientation of an acoustic source by a microphone array network, 2009 Acoustical Society of America, pp.3084-3094 (2009).

[12] Alberto Yoshihiro Nakano: Exploring Spatial Infomation For Distant Speech Recognition Under Real Environmental Conditions, Toyohashi University of Technology DOCTOR OF ENGINEERING (2010-03).

[13] SVM-Light Support Vector Machine, http://svmlight.joachims.org/

表 3. 方向別の前後判別正解率

方向	-180	-150	-120	-90	-60	-30	0	+30	+60	+90	+120	+150	+180	全方向
防音室	100.0	100.0	100.0	-	91.7	100.0	100.0	100.0	100.0	-	100.0	100.0	94.4	98.7
実環境	100.0	100.0	100.0	-	94.4	100.0	94.4	97.2	83.3	-	97.2	88.9	88.9	94.9

表 4. 距離推定の実験結果（50cm と 100cm の識別率）

学習・テスト方向	-180	-150	-120	-90	-60	-30	0	+30	+60	+90	+120	+150	テスト前方向平均	テスト全方向平均
	-150	-120	-90	-60	-30	0	+30	+60	+90	+120	+150	+180		
追加学習方向	+180											-180		
防音室	71.3	48.3	61.7	90.7	84.2	92.3	93.1	94.5	88.3	78.8	70.8	69.8	91.0	78.6
実環境	68.9	62.0	69.6	57.4	80.5	77.9	82.9	88.8	59.8	55.9	45.5	65.7	82.5	67.9
学習防音室,ﾃｽﾄ実環境	73.8	77.9	80.5	71.0	66.2	58.6	82.9	84.6	72.5	53.2	48.0	65.9	73.1	69.6