E-053

機械雑音混入音声中の音韻キュー探索による話者方向の同定

Speaker-direction Detection under Mechanical Noises based on Phoneme-cue Search

沼波 宰† 藤原 真志† 川端 豪†
Tsukasa Nunami Masashi Fujiwara Takeshi Kawabata

1．まえがき

ホームロボットやトイロボットに要求される人間との自然なコミュニケーションには、話しかけた人の方向を向いて返事をしたり、こちらへ歩いてきたりといった親密な動作が必要であり、このために話者の方向を同定する技術が必要となる。高度なロボットにおいては、視覚・聴覚の連携や多数のセンサの利用によってこれを実現するが[1,2]、トイロボットのように低価格の製品においては、最小限の費用でこれを行う必要がある。本報告では、ロボットに装着した 2 つのマイクのみを用いて、話者の方向を精度よく求める手法について述べる。

人間は両耳間の強度差や時間差を用いて音源方向を判定するが[3]、ロボットの場合はロボット自身が発する機械音が直接マイクに入るため、方向同定精度が上がらないことがある。藤原らは話者の口元に設置した近接マイクで得た音声成分を、ロボットの体に装着した左右 2 つのマイク入力信号の中から探し出し、その音声成分の時間差を求めることで、精度良く話者方向を同定することに成功した[4,5]。しかし、状況に応じては常に近接マイクを利用できるとは限らない。そこで本報告では、典型的な音声成分をあらかじめ用意しておき、近接マイク無しでも精度良く話者方向を判定する方法を提案する。

2．ロボットによる話者方向の同定

2.1 左右マイク信号の相互相関関数による時間差判定に基づく話者方向の同定（CC 法）

人間はある音源から音が発生した時に、その信号を両耳で捉えることによって、その音源の位置を同定することが可能である。その際に使われる特徴量の一つに ITD(Interaural Time Difference) がある[3]。

ITD とは、同じ音源から発生した音に対して、左耳と右耳の距離から生じる音の時間差のことを言う。人間の耳は、顔の正面を基準としてほぼ左右対称の位置にあるため、音源が真正面または真後ろ以外に位置を変えることで、左右の耳に到達する音には時間差が生じる。この時間差を方向同定の手掛かりとしている。ロボットにおいてこの ITD に類似した機構を実現するためには、例えばロボットの左右の肩にマイクを設置し、その入力信号の時間差により方向同定を行うことが考えられる。図 1 に示すように左右のマイク入力信号の相互相関(Cross Correlation)関数を計算し、そのピークを求めることにより時間差を計算する。以下、この方法を「CC 法」と記述する。

CC 法の深刻な問題としてロボットの内部雑音がある。ロボットの内部雑音(モーター音、機械音)はロボットに装着されたマイクに直接に大きな音量で入力され、しばしば人間の声よりも大きな音量になる。CC 法は、単純に左右マイク信号の相互相関関数を計算するので、左右内部雑音の相互相関が主要因となり、音声の到来方向を同定することができなくなる。

2.2 近接-遠隔マイクコンビネーションによる話者方向の同定（藤原法）

ロボット自身が発する機械音の混入という問題の対策として、藤原らは、近接-遠隔マイクコンビネーションによる話者方向の同定手法を提案した[2][3]。

この方法に基づく左右時間差による方向同定の方法を図 2 に示す。ロボットに設置されたマイク(話者からみて遠隔マイク)に加え、話者の口元に設置されたマイク(近接マイク)を利用する。遠隔マイクに入力される機械音が混入した信号の中から、近接マイクから得た音声信号の成分を探し出し、左右マイク中の音声成分の時間差を求めることによって、機械音の影響を抑制し、精度良い方向同定を行う。

図 2 に示すように、近接マイク信号と左右の遠隔マイク信号それぞれとの相互相関関数を計算し、ピークを探索することによって、近接マイク信号と「左」の遠隔マイク信号中の音声成分の時間差、及び近接マイク信号と「右」の遠隔マイク信号中の音声成分の時間差を求める。両者の差が左右マイク信号中の音声成分の時間差を表すことになる。この時間差に基づいて、話者方向の同定を行う。以下、この方法を「藤原法」と記述する。

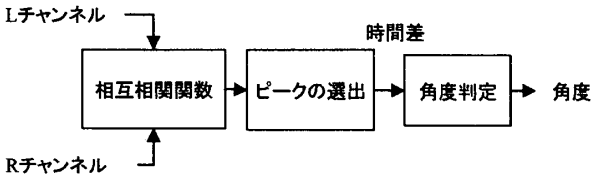

図 1：左右マイク信号の相互相関関数による時間差判定に基づく話者方向の同定（CC 法）

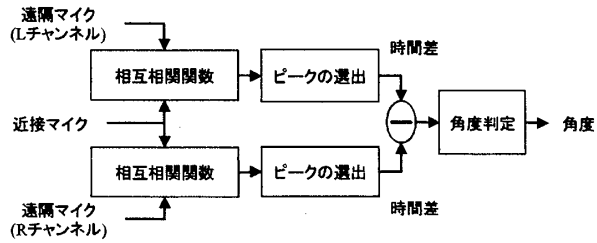

図 2：近接-遠隔マイクコンビネーションによる話者方向の同定（藤原法）

†関西学院大学 理工学部, KGU

2.3 音韻キュー探索による話者方向の同定（提案法）

藤原法は近接マイク信号の利用により、高精度の話者方向同定を可能にしたが、状況によっては近接マイクが使用できなかったり、近接マイクに何らかの雑音が混入し、うまく方向同定ができない場合もある。近接マイク信号がなくても、ある程度方向同定ができる手法を確立しておくことは、フォルトトレランス確立の観点からも意義があると考えられる。

藤原法では、近接マイクを利用し、その音声成分を遠隔マイク信号から探し出すことで方向同定精度を向上させていた。もし近接マイクの代わりとなり得る信号をあらかじめ記憶しておき、その成分を遠隔マイク信号の中から探し出すことができれば、藤原法に近い方向同定精度を達成できる可能性がある。

そこで本論文では、音韻キュー探索による話者位置の方向同定を提案する。例えば、典型的な音声の断片である母音や子音の音声信号を「音韻キュー」としてあらかじめ用意しておき、この音韻キュー成分を左右の遠隔マイク信号それぞれの中から探し、その時間差を求める。この作業を各音韻キューについて行い統合することで、近接マイクなしでも精度良く話者方向の同定を行うことを考える。

2.3.1 音韻キューの選択

人間が話す単語や文は、母音と子音の組み合わせによって生成される。この点から、音韻キューの候補としては、母音または子音を用いることが考えられる。

本研究の設定している実験条件ではロボットに装着されたマイクには機械音が大音量で混入している。子音は時間が短く、音量も小さめであるのでこの雑音に埋もれやすい。これに対し、母音は時間変化の広がりがあり、各種類それぞれの定常性を持ち、音量も大きい。

そこで今回は、機械音が混入した信号の中から、音声成分のみを探し出す手掛かり(音韻キュー)として、日本語の 5 母音を用いることにする。音韻キューの長さは、録音音声のサンプリング周波数 48kHz において 1024 点とする。またハミング窓を掛けることで、始点と終点付近に偶発的にできる強い特徴の影響を抑制する。

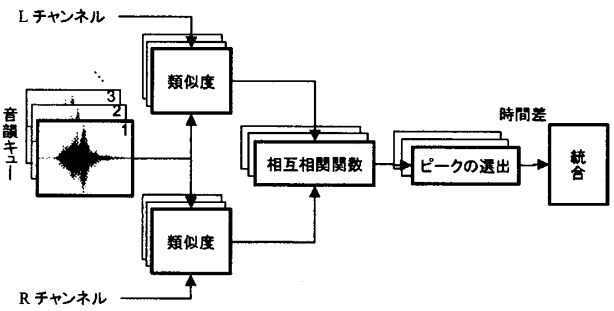

図 3：音韻キュー探索による話者方向の同定（提案法）

2.3.2 左右のマイク信号と音韻キューの類似度パターンの計算

提案法における処理の流れを図 3 に示す。

まず処理の第一段階として、機械音が混入した入力音声信号の中から音韻キュー成分を探し出す。この時、単純に両者の相互相関を計算すると入力音声信号の音量の大きいところで相互相関も大きくなるので、入力音声信号の音量で正規化することが必要である。この量はすでに単純な相互相関ではないので本論文ではこれを類似度と呼ぶことにする。

ある時刻 τ において切り出された分析窓長 M(=1024)の入力音声信号を $x(i)$ (i=0,1,2,...,M) 、音韻キューを $c(i)$ (i=0,1,2,...,M) とするとき、その時刻における類似度は次の式で計算される。

$$S(\tau)=\frac{\left[\sum_{i=0}^{M}x(i+\tau)\cdot c(i)\right]^2}{M^2\sum_{i=0}^{M}x(i+\tau)^2} \quad (1)$$

各時刻における類似度を計算し入力音声の長さの類似度パターンを作成する。

2.3.3 左右類似度パターンの時間差の計算

処理の第二段階は各音韻キューに対する左右類似度パターンの時間差を求めることである。各類似度パターンには入力音声信号中のどの時点にその音韻キューが含まれるかが反映されているので、左右類似度パターンの時間差を求めることで(機械音を抑制し)音声信号のみの時間差を計算できる。

以上の作業を、音韻キューの数だけ繰り返し、それぞれ時間差を求める。また、この作業に並行して CC 法による時間差も求めておく。

最後に、音韻キューの数だけある時間差と、CC 法による時間差を分散正規化距離により統合し、最終的な角度同定を行う。ここで、統合した分散正規化距離 D を次式に示す。

$$D=\frac{(\mu_{ITD}-x_{ITD})^2}{\sigma_{ITD}^2}+\sum_{j=1}^{N}\frac{(\mu_j-x_j)^2}{\sigma_j^2} \quad (2)$$

ただし、それぞれの角度において、単純なマイク入力信号による時間差の平均を μ_{ITD}、標準偏差を σ_{ITD}、新しい入力信号に対する時間差を x_{ITD} とし、音声キューを利用した場合における時間差の平均を μ_j、標準偏差を σ_j、新しい入力信号に対する時間差を x_j (j=1,2,...,N：N は音韻キューの数) とする。

分散正規化距離においては、ある特徴量の分散が大きくなり判定に有効でなくなると、自動的にその特徴量の判定への寄与度が下がる。このため CC 法による時間差を(2)式の第一項に含めておいても方向同定精度に対する悪影響はないと考えた。今後、より詳しく検討する必要がある。

3．評価実験

3.1 実験条件

実験には市販のトイロボットを利用した。ロボットの左右の肩それぞれに無指向性のマイクを装着した(図 4)。左右マイクの間隔は 15cm である。暗騒音 36dBA の防音室において、音韻の出現頻度を考慮した 50 単語をロボットに対して 5 方向(−60°，−30°，0°，30°，60°) (図 5)から、20 代前半の男性 1 名が発声した。発話者とロボットとの距離は 50cm 程度、比較実験のために話者の口元には近接マイクを設置した。ここで密接距離としたのは、本研究がトイロボットとの親密なコミュニケーションを目的としたためである。録音した音声のサンプリング周波数は 48kHz である。

音声を発声する際、ロボットを静止させた雑音無しの状態と、設置したマイクは移動させずロボットの腕を常に上下に動かし、機械音が混入する雑音有りの状態の 2 種類を上記 5 方向について行った。S/N 比は 10dB 程度である。

音声データの内、雑音無しで発声した 5 方向 50 単語、計 250 発話を学習データとして用いる。これは(2)式における各方向に対する時間差の平均と分散を求めるのに利用する。一方、雑音有りで発声した 5 方向 50 単語、計 250 発話を評価データとして用いる。前章で述べた 3 つの方式を用いて話者方向の同定を行う。

CC 法：左右マイク信号の相互相関関数による時間差判定に基づく話者方向の同定

藤原法：近接-遠隔マイクコンビネーションによる話者方向の同定

提案法：音韻キュー探索による話者方向の同定

ある発話のある方向からの入力に対して、正しい入射角度を判定できた場合のみを正答とする。

3.2 3 方式の性能比較

この節ではこれまで説明してきた 3 方式の方向同定の精度を比較する。表 1 に各方式の方向同定精度を示す。また表 2、表 3、表 4 にそれぞれの混同表を示す。

単純な左右の時間差に基づく CC 法の方向同定精度は 68.4%であった。表 2 を見ると多くのデータが 30° に誤判定されていることが分かる。これは機械雑音自体に対する左右時間差が、マイク位置やロボット内部のモーター位置の関係で、たまたまこの付近にピークを持つためである。

これに対し、遠隔マイクに入力される機械音が混入した信号の中から、近接マイクから得た音声信号の成分を探し出すことによって、左右マイク中の音声成分の時間差を求める藤原法においては方向同定精度は 80.8%に向上する。表 3 の混同表を見ても CC 法の場合に生じていた 30° への誤判定が緩和されていることが分かる。

音韻キュー探索に基づく提案法による方向同定精度は 74.8%であった。近接マイクを要する藤原法による方向同定精度 80.8%には及ばなかったが、CC 法の精度を大きく改善することができた。表 4 の混同表を見ると、全体の正答数では藤原法に劣るものの、誤判定においては提案法の方が正しい角度付近にまとまりが見られ、良好な方向同定特性を示している。

表 1：各方式の方向同定精度

方式	方向同定精度(%)
CC法	68.4
藤原法	80.8
提案法	74.8

表 2：CC 法による方向同定結果の混同表
(太枠箇所は発話方向が正しく同定された回数)

IN \ OUT	−60°	−30°	0°	30°	60°
−60°	**40**	1	3	5	1
−30°	7	**18**	**17**	4	4
0°		11	**19**	**20**	
30°				**50**	
60°				6	**44**

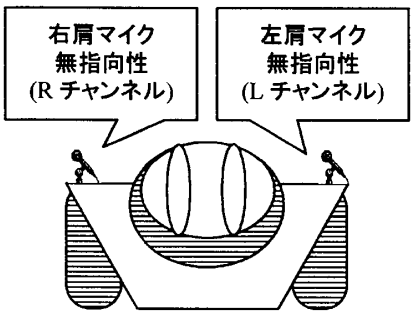

図 4：ロボットに装着したマイクの配置

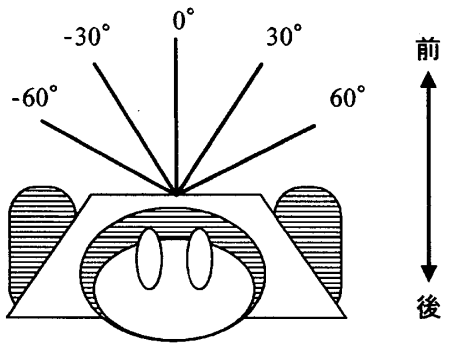

図 5：音声の入射角度

3.2.2 話者性の検討

先の実験では、必要な音韻キューを評価データと同じ話者の音声から切り出して利用した。しかし、ロボットに話しかける以前に、使用者の音声を録音し音韻キューを登録するのは不便である。この節では音声成分の探索に用いる音韻キューを評価データと異なる話者から作成した場合に、方向同定精度が保たれるかどうかを確認する。

表 5 に音韻キューの作成に用いる 3 人の話者を示す。話者 1 は評価データと同じ話者、話者 2 は評価データの話者と同じ性別同じ年齢の別の話者、話者 3 は同じ性別であるが年齢の異なる別の話者である。

表 5 の中にはこれらの話者の音声データを用いて音韻キューを作成した場合の方向同定精度を併せて記述してある。小規模な実験のため、この検討だけで提案法が音韻キューの話者性に依存しないとは言い切れないが、少なくとも異なる話者で音韻キューを作成した場合に、著しく精度が悪化しないことを確認した。

4. 結論

本報告では、音韻キューとして典型的な 5 母音の音声成分をあらかじめ用意しておき、近接マイク無しでも精度良く話者の方向同定を行う方法を検討した。単純な左右の時間差に基づく CC 法方向同定精度は 68.4%であった。音韻キュー探索に基づく提案法による方向同定精度は 74.8%であり、近接マイクを要する藤原法による方向同定精度 80.8%には及ばなかったが、CC 法の精度を大きく改善することができた。また、音韻キューの作成に異なる話者の音声データを用いても方向同定精度が著しく悪化しないことを確認した。

今回の実験はマイク 2 個のみを用い左右の方向同定のみを検討したが、今後はマイク数を増やし 360° の話者方向同定を検討したい。

文献

[1] 佐藤　幹, 杉山　照彦, “パーソナルロボット PaPeRo における音声インターフェイス”, 日本音響学会誌, 62(3) (2006) 173-181

[2] 浅野　太, “ロボットにおける音源位置推定”, 日本音響学会誌, 63(1) (2007) 41-46

[3] Jeffress, L.: A place theory of sound localization. J. Comp. Physiol Psychol. 41 (1948)35-39

[4] Kawabata, T., Fujiwara, M., Shibutani, T.: Detection of Speaker Direction based on the On-and-Off Microphone Conbination for Entertainment Robots. Entertainment Computing - ICEC 2005 (2005) 248-255

[5] 藤原　真志, 川端　豪, “近接-遠隔マイクコンビネーションによる全方位型話者方向同定”, 信学技法, (2006-06) 13-18

表 3：藤原法による方向同定結果の混同表
(太枠箇所は発話方向が正しく同定された回数)

IN \ OUT	−60°	−30°	0°	30°	60°
−60°	**43**	**2**	**2**	**2**	**1**
−30°		**40**	**7**	**3**	
0°	**6**	**1**	**35**	**7**	**1**
30°	**1**	**1**	**5**	**41**	**2**
60°	**1**		**4**	**2**	**43**

表 4：提案法による方向同定結果の混同表
(太枠箇所は発話方向が正しく同定された回数)

IN \ OUT	−60°	−30°	0°	30°	60°
−60°	**48**	**1**	**1**		
−30°	**12**	**29**	**9**		
0°		**5**	**40**	**2**	**3**
30°			**25**	**22**	**3**
60°	**1**		**1**		**48**

表 5：音韻キューを作成した話者の説明

音韻キュー作成用話者	評価データとの関係	性別	年齢	方向同定精度(%)
1	同じ話者	男	23	74.8
2	無関係	男	23	74.0
3	無関係	男	50	78.4