2007－SLP－67 (11)
2007／7／21

ロボットに装着された３個のマイクによる話者方向の同定

沼波 宰　　藤原 真志　　川端 豪
関西学院大学　理工学部　情報科学研究科
email:{bhz80890, fujihara, kawabata}@ksc.kwansei.ac.jp

概要：音韻キュー探索に基づく話者方向同定の新しい手法を提案する。まず、モーター音や機械音からなる直接雑音が混入した音声中から複数の音韻の成分を探索し、次に、左右チャンネルの信号中から探索された母音部分の時間差を求める。単純な相互相関関数は直接雑音の影響を受けてしまうので、本手法では各チャンネルと音韻キューとの類似度パターンを計算し、類似度パターン間の相互相関関数によってチャンネル間の時間差を求める。3つのマイクをロボットの左右の肩と胸に設置することで三角形を形成し、全方向の方向同定を行う。

Speaker-direction Detection with 3 Microphones under Mechanical Noises based on Phoneme-cue Search

Tsukasa Nunami, Masashi Fujiwara, and Takeshi Kawabata
Kwansei Gakuin University

Abstract: This paper proposes a new method for detecting the speaker-direction based on a phoneme cue search approach. The system searches for the spectral elements of several phonemes from noisy speech with direct motor and mechanical noises. For example, the time delay of vowel parts between the left and right channels indicates the time delay of the left and right channels. Because a simple cross correlation method is disturbed by the direct noises, our system calculates the cross correlation of the vowel elements of target channels. Three microphones, located at robot shoulders and a chest, make a triangle and enable all directional speaker-direction detection.

１．　まえがき

近年盛んに人間型ロボットの研究が行われるようになってきた。自立二足歩行をする高度なレベルのものからペット的なトイロボットまで多種多様なレベルのものが存在する。ホームロボットやトイロボットを考えるとき、人間との自然なコミュニケーションは欠かせない要素だと思われる。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望まれる。このためには話者の方向を同定する技術が必要である。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実現する必要がある。

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[3]が、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。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 3つのマイク信号を利用した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、ある程度の音源位置を同定することが可能である。その際に使うと考えられている物理現象に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であり、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。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、方向同定の手掛かりとして用いられる。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。しかしこの方法は、ロボットの前方と後方で左右のマイク入力信号の時間差が同じとなる点が存在してしまうために前後の誤判定が起こりやすく、全方位の方向同定を高精度で行うことができない。

この問題の解決策としてHuangの研究がある[6]。この方法は、3つのマイクを三角形になるように配置することで、2つのマイクでは困難であった前後判定を可能としている。

ある2つのチャンネル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方法を図1に示す。この作業を三角形の各辺について、すなわち左右の肩マイク、左肩マイクと胸マイク、右肩マイクと胸マイクの3通りについて行い、それぞれ時間差を計算する。3つの時間差を式(1)に示す分散正規化距離尺度で統合することにより方向同定を行う。

$$D_{CC} = \frac{(\mu_{ITD} - x_{ITD})^2}{\sigma_{ITD}^2} \qquad (1)$$

ここで、それぞれの角度において、三角形のある一辺のマイク入力信号の時間差の平均を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}とする。以下、この方法を「CC法」と記述する。

CC法による方向同定精度を悪化させる1つの原因として、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。この場合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[4,5]。この方法に基づく2つのチャンネルのマイク信号の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すことができる。

例えば、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。この作業を、左右の肩の遠隔マイク、左肩と胸の遠隔マイク、右肩と胸の遠隔マイク、計3通りについて行い、それぞれの時間差を、分散を正規化した距離尺度により統合し、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。

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。以下、詳しく述べていく。

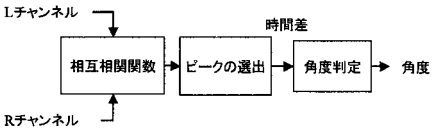

図 1：左右マイク信号の相互相関関数による時間差判定に基づく話者方向の同定（CC 法）

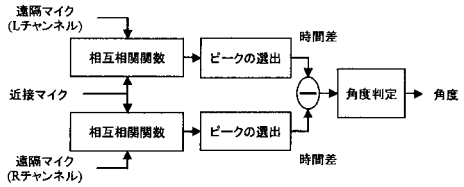

図 2：近接-遠隔マイクコンビネーションによる話者方向の同定（藤原法）

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。

2.3.2 マイク入力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、例えば左右のマイク入力信号に基づく処理の流れを図 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 （i=0,1,2,...,M）、音韻キューを （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。

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

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 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それぞれの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、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。

左右の肩マイク、左肩マイクと胸マイク、右肩マイクと胸マイクの 3 つの組み合わせ各々について、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ておく。また後に述べる理由により CC 法による時間差も同時に求めておく。

最後に、3 つのマイクの組み合わせ、計 3 通りについて、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 を次式に示す。

$$D=\frac{(\mu_{ITD}-x_{ITD})^2}{\sigma_{ITD}^2}+\sum_{j=1}^{N}\frac{(\mu_j-x_j)^2}{\sigma_j^2} \qquad (3)$$

ここで、それぞれの角度において、三角形のある一辺におけるCC法に基づく時間差の平均を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は音声キューの数) とする。

分散正規化距離には優れた特徴がある。ある特徴量が大きな分散を持ちある判定に対して有効で無い場合、分母が大きくなることによってその特徴量の判定に対する寄与度が自動的に下がる。故に、判定に対して有効かどうか不明な特徴量であってもとりあえず分散正規化距離の要素として入れておいても害は少ない。逆に何らかの条件下でその特徴量が有効である可能性があるならばむしろ積極的に要素として入れておく方が良い。この観点から(3)式の分散正規化距離の計算にはCC法によって計算される時間差の項を残した。

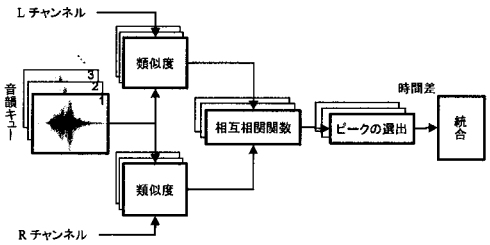

図3：音韻キュー探索による話者方向の同定（提案法）

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、胸それぞれに無指向性のマイクを装着した(図4)。それぞれのマイクの間隔は15cmである。暗騒音が36dBAである防音室において、音韻の出現頻度を考慮した50単語をロボットに対して12方向(0°，30°，60°，90°，120°，150°，180°，210°，240°，270°，300°，330°) (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。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12方向について行った。S/N比は10dB程度である。

雑音無しで発声した12方向50単語、計600発話を学習データとして用いる。これは式(3)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12方向50単語、計600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

ある発話のある入射角度での入力に対して、正しい入射角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。

CC法の方向同定精度は41.7%であった。表2を見ると誤り先が120°、150°、270°となっている誤判定が多く観察された。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はないかと考えている。

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74.8%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後方3方向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76.1%であった。近接マイクを利用する藤原法と比べて良くなっている所も悪くなっている所もあるが全体として誤判定が減少し、藤原法を超える方向同定精度を実現することができた。

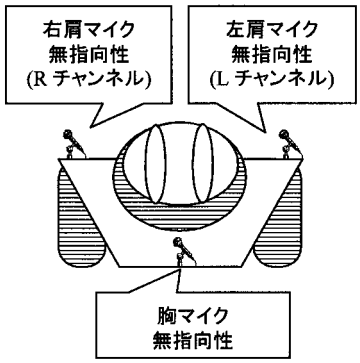

図 4：ロボットに装着したマイクの設置図

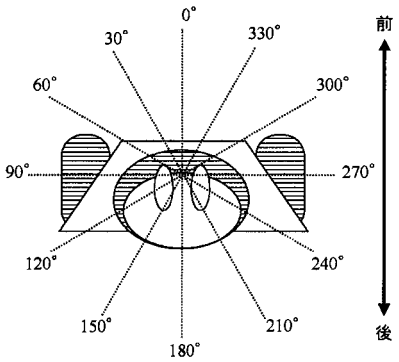

図 5：音声の入射角度

表 1：各方式の方向同定精度

方式	方向同定精度(%)
CC法	41.7
藤原法	74.8
提案法	76.1

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	0	30	60	90	120	150	180	210	240	270	300	330
0	19	17	1		1	1				6		5
30	6	32	1		1	4				5		
60		6	31		6	7						
90	1	4	2	10	29	3				1		
120	2		2	1	29	15				1		
150				2	16	26				5		1
180					3	5	18	1		7		16
210			1			9	3	8	5	9	9	6
240	1					4	5	10	9	4	15	2
270		1	1		2	2	2	4	3	33	1	1
300	4	3	1		2			1	2	17	19	1
330	11	1	1	1	1	5	7	1		2	4	16

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	0	30	60	90	120	150	180	210	240	270	300	330
0	38	5								3	3	1
30	8	35	1				5					
60	3	2	42			1	1	1				
90		4	1	38	7							
120	1	2	1	3	39	3						1
150	3	3		2	2	36	1			3		
180	2		1	1		1	41	3		1		
210	5	1					3	36		5		
240	3	1				2	2		26	14	2	
270	1	1				1			1	45	1	
300	4	2								6	38	
330	9	3					2			1		35

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	0	30	60	90	120	150	180	210	240	270	300	330
0	47	1	1		1							
30	31	17	1									1
60	1		46	3								
90	1			40	9							
120				1	42	7						
150					6	41	3					
180						4	46					
210							6	34	4	4	2	
240								6	35	9		
270								1	4	33	12	
300									1	5	44	
330	12									6	32	

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。結果を比べてみると、評価データと同一話者の音韻キューを使用した時よりも、評価データとは異なる話者の音韻キューを使用した時の方が方向同定精度が低い。今後、この問題を解決するために、どのような音声に対しても高精度で方向同定ができるような音韻キューの検討をしていきたい。

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法による方向同定精度は41.7%であった。また、近接マイクを要する藤原法による方向同定精度は74.8%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これを更に超える76.1%を達成した。このように近接マイクを用いなくても音韻キューの成分を機械音が混入した信号中から探し出しその時間差を計算することで、CC法の精度を大きく改善することができた。今後、話者に依存しない音韻キューの作成について検討していく必要がある。

表5：各方式の方向同定精度

音韻キュー作成用の話者	説明	性別	年齢	方向同定精度(%)
1	評価データと同一の話者	男	23	76.1
2	評価データとは無関係	男	23	70.5
3	評価データとは無関係	男	50	72.8

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

[6] J.Huang, N. Ohnishi, and N.Sugie. Building ears for robots: sound localization and separation. Artificial Life and Robotics, 1(4) (1997)157-163