

ロボットに装着されたマイク信号中からの 音韻キュー探索による話者方向の同定

沼波 幸 川端 豪

関西学院大学 理工学部 情報科学研究科

email:{bhz80890, 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。三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。それぞれのチャンネルに対し、機械音からなる直接雑音が混入した音声の中から複数の音韻の成分を探索し、その探索された母音部分の時間差を三角形の各辺について求め、統合することで角度判定を行う。本報告では、まず異なる話者の音韻キューを使用した場合の性能劣化を検証する。その後、本人を含めた複数話者の音韻キューを併用したマルチテンプレート法の性能を示す。

Speaker-direction Detection under Mechanical Noises based on Multi-speaker Phoneme-cue Search Tsukasa Nunami, Takeshi Kawabata Kwansei Gakuin University

Abstract: This paper describes an improvement of the speaker-direction detection method based on the multi-speaker phoneme-cue search. Three microphones, located at robot shoulders and a chest, make a triangle and enable all directional speaker-direction detection. The system searches for the spectral elements of several vowels with the three-channel noisy speech with direct motor and mechanical noises. The time delays among these three channels indicate the speaker direction. First, this paper shows the degradation of direction-detection accuracy in the case that another person's vowels are used for phoneme cues. After that, we show the effectiveness of the multi-speaker phoneme-cue approach.

1. まえがき

ロボット研究の発展には目覚ましいものがあり、近い将来、ホームロボットとして日常の生活にロボットが活躍する時代も遠くないと思われる。その実現のためにロボットに要求される課題の一つとして「人間との音声によるコミュニケーション」がある。この課題は、視覚、聴覚、対話など様々な技術要素の集合であるが、その中の一つに話者方向の同定がある。

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[1]が、ロボットの場合はモーターなどの機械

雑音が直接マイクに混入するため、単純な強度差や時間差では方向同定の精度が上がらない。

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[2,3]。この手法は強力であるが、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、フォルトトランス確立の観点からも意義があると考えられる。

前報[4]において、近接(口元)マイクの入力信号の代わりとなる音声成分をあらかじめ記憶しておき、機械雑音の混入した遠隔(ロボットに装着された)マイクの信号の中から、音声キューの成分を探し出し、時間差を求めることで、話者方向の同定を行う方法を提案した。この時、音韻キューとしては本人が前もって発声した母音が有効であることも示した。

本報告では、まず音韻キューとして本人以外の母音を用いた時に、上記手法の方向同定精度がどのくらい保たれるかを観察する。また複数の話者の母音を同時に音韻キューに利用することによって方向同定精度がどのように変化するかを検討する。

2. 音韻キュー探索による話者方向の同定

まず、前報において提案した音韻キュー探索による話者方向同定の処理手順について述べる。実際には全方向の同定を行うために三角形に配置したマイクロホンを用いるが[5]、ここでは簡単のためにロボットの左肩マイクと右肩マイクの二本のマイクで説明する。図1に処理手順を示す。まず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く。次に各音韻キューについて、その音声成分とロボットの左右マイク信号との類似度パターンを計算する。これによって機械雑音混入音声の中から音声成分だけを取り出す。こうして計算された2つの類似度パターンの相互相関を計算し、左右マイク信号の時間差を求める。最後に音韻キューの数だけ求められた時間差を分散正規化距離尺度により統合することで精度良く話者方向の同定を行う。以下、これらの詳細を順に示す。

2.1 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。

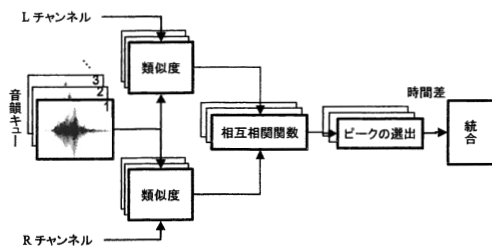


図1：音韻キュー探索による話者方向の同定

音韻キュー候補として検討した母音と子音の種類を表1に示す。ここで、/*_start/と表現されたものは、無音区間からの立ち上がり母音を意味する。このキューを候補に入れたのは、立ち上がりの瞬間に強い特徴が現れ、音声成分探索の手掛かりとして有効であると考えたためである。本手法において効率よく精度を向上させるためには、表1における数多くの母音や子音の中から、実際に機械雑音混入音声声中において、その音声成分を探し出すのに有効な音韻キューを選択する必要がある。そこで、図1の処理手順において単一の音韻キューを用いて方向同定精度を求め、有効な音韻キューの選択を行った。全般的に子音よりも母音の方が精度が高いことが分かった。

また、機械雑音混入音声と音韻キューの類似度パターンを観察してみると、子音は雑音と反応してはばばらに反応しているのに対し、母音は機械音に影響されことなく母音部分で反応していることが分かる。これは、子音は時間が短く、音量も小さめであるので雑音に埋もれやすいが、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つためであると考えられる。

以上の点を踏まえ、機械音が混入した信号の中から音声成分を探し出す手掛かりとなる音韻キューとして、日本語の5母音を用いることにした。その種類を表2に示す。

なお、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点とし、またハニ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している。

表1：音韻キュー候補の母音と子音

a	i	u
e	o	a_start
i_start	u_start	o_start
ch	d	j
k	s	sh
t	z	

表2：実際に使用する音韻キューの種類

a_start	i_start	u	e	o_start
---------	---------	---	---	---------

2.2 マイク入力信号と音韻キューの類似度パターンの計算

ここでは、前報[4]で導入した類似度計算法について説明する。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ために、式(1)の類似度を計算する。分子は入力音声と音韻キューの畳み込み和を計算しており、分母は入力音声の大きさを正規化する。この式は結局、相互相関関数を入力音声の音量で正規化したものと考えられる。

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

ここで、ある時刻 τ において切り出された分析窓長 $M(=1024)$ の入力音声波形を $x(i)$ ($i=0,1,2,\dots,M$)、音韻キュー波形を $c(i)$ ($i=0,1,2,\dots,M$) とする。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している。

このように前報では、時間領域の計算によって入力音声波形と音韻キューとの類似度パターンを計算している。しかし、今後話者に依存しない方向同定手法を確立することを想定し、時間領域の計算よりも話者性が良く現れるスペクトル領域で同等の計算を行う方が後に見通しが良くなると考え、スペクトル領域で類似度を計算することにした。この場合の計算式を式(2)に示す。

$$S_{\tau}(f) = \frac{X_{\tau}(f)C^{*}(f)}{M \sum_{\tau} X_{\tau}(f)} \quad (2)$$

ここで、ある時刻 τ において切り出された分析窓長 $M(=1024)$ の入力音声信号の FFT(Fast Fourier Transform) を $X_{\tau}(f)$ 、音韻キューの FFT を $C(f)$ (C^{*} は共役複素数、 f は周波数) とする。FFT の点数は $M(=1024)$ 点である。

式(2)を使用した類似度パターン生成の手順として、まず時刻 τ で切り出した M 点の信号に後の FFT を考慮しハニング窓を掛ける。次に切り出した信号と使用する音韻キューそれぞれを FFT し、式(2)の計算を行う。こうして計算された類似度を IFFT することで時刻 τ から M 点分の類似度パターンができる。 τ を $M/2(=512)$ 点ずつずらして同じ計算を行い、それら類似

度を加算していくことで、全体の類似度パターンを作成する。

このようにして、全ての音韻キューと全ての入力チャンネルの組み合わせに対し、類似度パターンを計算する。

2.3 類似度パターンの時間差の計算

ある音韻キューと各入力チャンネルの類似度パターンを用いてそのチャンネル間の時間差を計算する。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、類似度パターンの出現する位置の時間差を求めることで(機械音を抑制し)音声信号の時間差を計算できる。

ここでも前項と同様にスペクトル領域にて時間差計算を行う。2つの信号の時間差を求める相互相関関数のフーリエ変換はクロススペクトルで定義される。

クロススペクトルによる類似度パターンの時間差を求める手順として、まず2つの類似度パターンから時刻 τ で $M(=1024)$ 点の分析窓を切り出し、ハニング窓を掛ける。次に、それぞれの分析窓信号を FFT し、クロススペクトルを計算する。この作業を、 τ を $M/2(=512)$ 点ずつずらしながら類似度パターンの終点に至るまで繰り返しつつ加算していく。最後に加算した回数でクロススペクトルを平均化し、その IFFT を行うことで相互相関関数を得る。

後述するようにロボットの左右の肩のマイクロホンの間隔は約 15cm である。日常温度の音速を 340m/s とすると、左右のマイクに生ずる最大時間差は約 0.4ms である。標準化周波数を 48kHz とするとこの時間差は 20 点に相当する。故に、相互相関関数のピーク検出時に範囲を制限することで、計算処理の効率を向上させ、誤判定を減らす。

2.4 各音韻キューに対する時間差の統合

処理の最終段階として、音韻キュー数×チャンネルの組み合わせ数だけ計算された全ての時間差を統合し、音声到来方向の判定を行う。統合は式(3)に示す分散正規化距離を用いて行う。各入射角度の学習データについて時間差に関する平均と分散を求めておき式(3)を適用する。最も距離が小さくなる角度を音声到来方向と判定する。

$$D = \sum_{j=1}^N \frac{(\mu_j - x_j)^2}{\sigma_j^2} \quad (3)$$

ここで、それぞれの角度において、音声キューを利用して求められた時間差の平均を 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 ($j=1,2,\dots,N$: N は音声キューの数) とする。

3. 評価実験

本報告では、まず機械雑音混入音声から音声成分だけを取り出す音韻キューとして、話者本人の母音を用いる場合と、異なる話者の母音を用いる場合の性能を比較する。次に、複数話者の音韻キューを統合利用したマルチテンプレート法の性能評価を行う。両方の実験に共通した実験条件を以下に示す。

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 2, 3)。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 4)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、録音した音声のサンプリング周波数は 48kHz である。

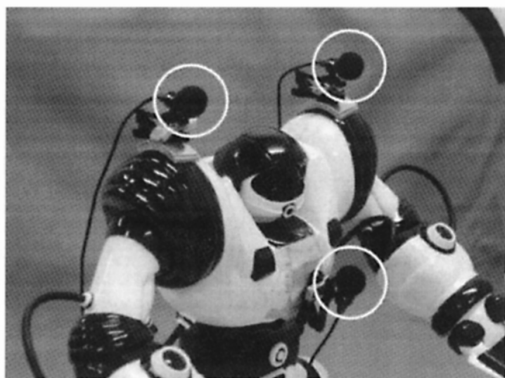


図 2：ロボットに装着したマイク(写真)

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。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。

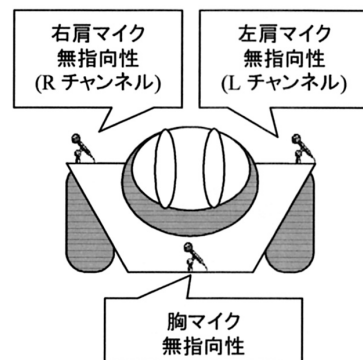


図 3：ロボットに装着したマイクの設置図

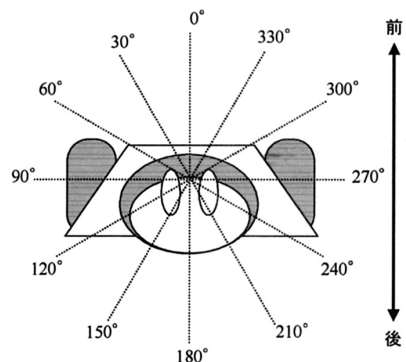


図 4：音声の入射角度

3.2 異なる話者の母音を音韻キューとして用いた場合の方向同定精度

この節では、音韻キュー探索による話者方向同定において、評価データと同一の話者の母音を音韻キューとして用いた場合と、異なる話者の母音を音韻キューとして用いた場合の方向同定精度を比較する。

音韻キュー作成用の話者 1, 2, 3 の内、話者 1 は評価データ本人、話者 2, 3 は異なる話者である。それぞれに対する方向同定精度を表 3 に示す。また各場合の混同表を表 4, 表 5, 表 6 に示す。

話者 1(本人)の音韻キューを利用した方向同定精度は 73.7%であった。表 4 を見ると、他の角度に比べて誤り先が 0°、120°、270° となっている誤判定がやや多い。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ち、その影響を受けたのではないかと考えている。

話者 2 の音韻キューを利用した方向同定精度は 68%であり、話者 1 の音韻キュー利用時よりも精度が大きく低下する。表 5 の混同表を見ると、ある角度においては話者 1 の時よりも良い精度を得られているが、それ以外のほとんどの角度では正答数が下がり、著しい精度低下が見られる。これは、話者 2 の声の特徴が話者 1 の声の特徴と大きく異なり、不安定な結果に繋がったと考えられる。

話者 3 の音韻キューを利用した方向同定精度は 72.5%であり、話者 1 のキュー利用時の精度と比べて著しい低下は見られなかった。表 6 を見ると、270° への大きな誤判定が見られるものの、それ以外は話者 1 の場合の表 4 とほとんど同じ傾向が見られる。これは話者 2 の時とは逆に、話者 3 の声の特徴が話者 1 の声の特徴と類似していたことが原因として考えられる。

表 3：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	73.7
2	評価データとは無関係	男	23	68
3	評価データとは無関係	男	23	72.5

表 4：話者 1 の音韻キューによる同定結果の混同表
(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	41	1	2	1								5
30	31	14	5									
60	1		40	9								
90				39	11							
120				5	42	3						
150					11	31	8					
180						3	44	3				
210							2	30	14	4		
240								3	37	10		
270										43	7	
300	1									6	41	2
330	6										4	40

表 5：話者 2 の音韻キューによる同定結果の混同表
(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	42	1		1							5	1
30	30	12	4	1							1	2
60	5		40	4	1							
90			4	29	17							
120					46	4						
150					5	28	17					
180						2	48					
210							3	36	4	7		
240								6	23	20	1	
270										36	14	
300										2	48	
330	5										25	20

表 6：話者 3 の音韻キューによる同定結果の混同表
(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	44		1	1							1	3
30	19	30										1
60	2		43	5								
90	1			46	3							
120	1			12	33	4						
150					12	30	6					2
180						1	45	2			2	
210								26	13	11		
240								6	13	31		
270	1								1	1	46	1
300	1									10	38	1
330	7										2	41

3.3 複数話者の全音韻キューを利用したマルチテンプレート法の検討

今までに説明した方法は、ある一人の話者に対して、ある一人の話者の音韻キューを利用した場合を想定している。前項からも分かるように、話者本人以外から採取した音韻キューを利用すると方向同定精度が低下してしまう。これは、それぞれの話者の母音のスペクトルに差異が見られるからであると考えられる。今後、話者性に依存しない方向同定手法へと改良していく予定であるが、今回はこのための一検討として、ある一人の話者に対して、その人を含む複数の話者から採取した音韻キュー全てを利用したマルチテンプレート方式を試みた。結果を表 7 に示す。またこの方式の混同表を表 8 に示す。

マルチテンプレート方式の方向同定精度は 76.2% であり、本人の音韻キューを利用した場合の方向同定精度よりも向上させることができた。表 2、表 3、表 4 それぞれと表 5 を比較すると、前者に見られる 240° から 270° への誤判定が後者では更に大きくなり、240° の方向同定精度が著しく低下している。これは 3 者それぞれの音韻キューを利用した時の傾向が、全て統合することでより大きい形で表れたものだと考えられる。

表 7: マルチテンプレート音韻キューを用いた場合の方向同定精度

音韻キュー作成用の話者	説明	方向同定精度 (%)
1+2+3	話者3人によるマルチテンプレート法	76.2

表 8: 全 3 人の音韻キューによる同定結果の混同表 (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	43		1	1							1	4
30	29	17	3									1
60	2		44	4								
90				43	7							
120				3	44	3						
150				8	33	9						
180					2	46	2					
210						1	35	9	5			
240							6	23	21			
270									46	4		
300									7	43		
330	5										5	40

4. 結論

ロボットが日常生活に溶け込み、自然に人間と共存するためには音声による対話は不可欠な要素である。本報告ではそのための要素技術の一つである話者方向の同定について検討した。以前に報告した人間の口元マイクを利用する方法をヒントに、口元マイクが不要な手法として「音韻キュー」を記憶しておく方式を検討している。

今回は、本人も含めた複数話者の全音韻キューを利用することによる話者方向の同定手法を提案し、その結果、本人だけの音韻キューを利用した場合より精度を向上させることができた。

異なる話者の音韻キューを更に加えることでより精度が向上することが期待できるが、現在の枠組みでは、話者が N 人になれば計算量も N 倍となり、リアルタイム性が失われてしまう。今後、計算を最適化したり、音韻キューを構造化することで計算量を大きく増加させずに不特定話者の音声を取り扱えるような手法に発展させたい。

文献

- [1] Jeffress, L.: A place theory of sound localization. J. Comp. Physiol Psychol. 41 (1948)35-39
- [2] Kawabata, T., Fujiwara, M., Shibutani, T.: Detection of Speaker Direction based on the On-and-Off Microphone Combination for Entertainment Robots. Entertainment Computing - ICEC 2005 (2005) 248-255
- [3] 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
- [4] 沼波 宰, 藤原 真志, 川端 豪, “ロボットの装着された 3 個のマイクによる話者方向の同定”, SLP, (2007-07) 57-62
- [5] J.Huang, N. Ohnishi, and N.Sugie. Building ears for robots: sound localization and separation. Artificial Life and Robotics, 1(4) (1997)157-163