人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A802

AIチャレンジ研究会 (第28回)

Proceedings of the 28th Meeting of Special Interest Group on AI Challenges

CONTENTS

日 時 2008年11月18日 場 所 京都大学 百周年時計台記念館 国際交流ホール
Kyoto University Clock Tower Centennial Hall, Kyoto, Nov. 18, 2008

社団法人 人工知能学会
Japanese Society for Artificial Intelligence

共催 京都大学グローバル COE プログラム
「知識循環社会のための情報学教育研究拠点」
Kyoto University Global COE Program:
Informatics Education and Research Center for Knowledge-Circulating Society

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A802-1 (11/18)

音の時間的な連続，不連続に関わる錯覚現象
Auditory Illusions Related to Temporal Continuity and Discontinuity

○中島 祥好 (九州大学芸術工学研究院)
* Yoshitaka NAKAJIMA (Kyushu Univ.)

nakajima@design.kyushu-u.ac.jp

Abstract— This presentation is a theoretical attempt to understand three auditory illusions related to the temporal continuity and discontinuity of sounds. In 1) the auditory continuity illusion, a long tone of 1-2 s interrupted by a short noise of 0.1-0.3 s, for example, is often perceived as continuous. In 2) the gap transfer illusion, typically a glide of 1.5 s or above with a temporal gap of about 0.1-0.3 s in the middle and a shorter continuous glide of about 0.4-0.6 s cross each other at their central positions; the gap in the longer glide is perceived as if it were in the shorter glide. In 3) the illusory gap unification (the illusory auditory completion), typically a glide of 1.5 s or above and a shorter glide of about 0.4-0.6 s cross each other, sharing a gap of 0.05 s or below; this gap is perceived more clearly, or only, in the shorter glide. New versions of these illusions in which harmonic glide tones have been employed are demonstrated. They are more persuasive than our previous demonstrations, in which single-component tones were employed mainly. A model seems to explain the illusions, and it is based on the assumption that the auditory system detects and combines onsets and offsets (terminations) of sounds as independent components.

1. はじめに

聴覚システムの重要な働きの一つとして，時間的に一つながりに聴こえる音の系列，すなわち音脈に含まれる音の数を決定するということがある。音脈は通常，音と空白時間との連鎖として捉えることができる。時間のなかでの音と音との境界，あるいは音と空白との境界のような，知覚内容の上で不連続である箇所がどのように決定されるかを解明することが，ヒトの聴覚コミュニケーションの仕組みを理解するうえで不可欠である。本稿においては，この問題をとり扱ううえで鍵となる聴覚上の錯覚現象(錯聴) を三つ紹介する。そのうちの二つについては，筆者も発見者に名を連ねているので，まだ雑誌論文として発表するに至っていない事柄についても，適宜記しておきたい。 (文章を読んだだけでは感じをつかみにくい箇所については，聴覚デモンストレーションを予定している。希望があればそのファイルを提供することもできる。)

ところで，聴覚における錯覚現象 auditory illusion を示すために「錯聴」という用語を用いることを提唱したい。その理由は，視覚における錯覚現象 optical illusion – この英語は不適切であるがここでは深入りしない – を示す「錯視」という用語が大いに普及しているため，日本語の文脈において，「錯覚」といえば「錯視」のことであると思われる傾向が強いと考えられるからである。実際に，専門外の方と話をする場合，聴覚にも錯覚があると言うと驚かれる場合が多く，ある程度勉強されている方も，マガーク効果などの聴覚と視覚とにまたがる例をまず挙げられる場合が多い。このような状況を改善するには「錯聴」という用語を普及させることも有効ではないかと考える次第である。

聴覚心理学の研究者が錯聴に興味を引かれるのは，聴覚システムの働きを理解するうえで，システムがうまく働かない事例が役立つと考えるからであり，視覚心理学の研究者が錯視を取りあげるのと同じような理由による。しかし錯聴は，錯視のように日常生活の中に偶然見つかるということが殆どありえない。わざわざ音を作ったり，分析したりしたときに初めて見つかるものである。戦前までは，技術的な制約のために聴覚研究者が作ることのできる音のパターンが限られており，結合音や分割時間の過大評価など，ごく僅かな錯聴が報告されているに過ぎない。ちなみに，このころまでに初等的な心理学の教科書に登場する錯視は，おそらく半分以上が出揃っている。ところが，20 世紀の末 (AMIGA、IBM 互換機、Macintosh が覇権を競っていたころ)から，研究者個人がコンピューターを占有して気軽に様々な音のパターンを作ったり，分析したりすることができるようになった。21 世紀初頭の現在においては，学生がアルバイトで貯めた程度のお金で，必要な装置一式が手に入る (測定器は別であるが) 。将来の研究の基礎となるような錯聴を発見したければ，現在は絶好の時期であり，適切な心がけさえあればアマチュア研究者が仲間入りすることも可能である。

2. 時間的連続性に関する錯聴

連続聴錯覚 The auditory continuity illusion

この現象は，おそらく最もよく知られた錯聴であり，さまざまな変形パターンも考案されている (Miller & Licklider, 1950; Ciocca, 1987; Bregman, 1990; Warren, 2008; Riecke, van Opstal, & Formisano, 2008) 。典型例を一つ挙げるならば，1000 Hz，1～3 s の純音の中央に，0.1～0.3 s の時間的空隙を作ると，この空隙は明瞭に聴きとられる。ところが，この空隙を，充分に強い– 例えば純音よりも 20-30 dB 強い – 500-2000 Hz の帯域雑音で埋めると，純音が再び連続

したものとして知覚されることが多い (Figure 1; 図示した刺激パターンは、Figure 6 を除いていずれも筆者のホームページで聴くことができる。)。

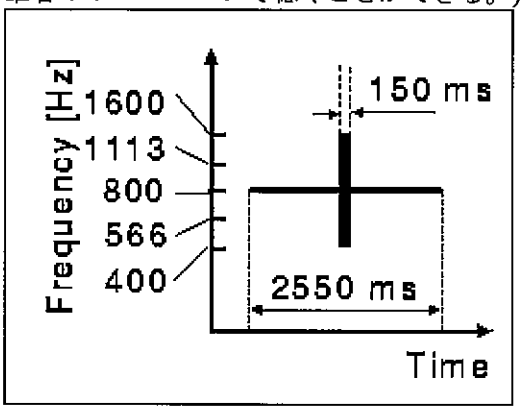


Figure 1. A stimulus pattern that is likely to cause the auditory continuity illusion. A temporal gap of 0.15 s in the pure tone is replaced with an intense noise. The tone is perceived as continuous if the noise is intense enough. (Most of the illustrated stimulus patterns in this paper are available as wave files on the author's website.)

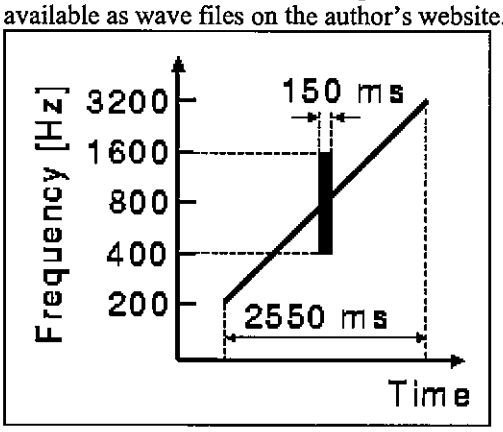


Figure 2. A stimulus pattern with a glide tone which is subject to the auditory continuity illusion.

純音のかわりに周波数変化音を用いても、この錯聴は生ずる (Ciocca & Bregman, 1987; Figure 2)。

この錯聴が生ずるために，二つの条件が必要であると考えられることが多い。まず，雑音の直前，直後の純音の切れ目は，雑音と重なっているか極めて近接するかのいずれかとなり，知覚のうえでかき消されることが，必要である。次に，雑音の純音付近の帯域に空隙の導入によって失われた音エネルギーを供給するくらいの音エネルギーの存在することが必要である。仮に純音が空隙を有せず連続していたとしても，雑音が空隙のあった部分の純音をマスキングによってかき消してしまうことが，雑音が連続聴を生ずるための要件であると考えられる場合が多い。連続聴錯覚そのものは広く知られており，その存在を疑うものはいないが，どのような条件においてこの錯覚が生ずるかについては，精神物理学的なデータがさらに蓄積されることが必要であろう。音の大きさの等感曲線を求めることや，同時マスキングの生起条件を決定することが，聴覚の基本的な仕組みを解明するうえで重要であったのと同じように，連続聴錯覚の生起条件に関して，系統だった実験データを提供することの意義は大きいであろう。

これまでの研究から確実に言えることの一つは，時間的空隙を埋める音が，もともと空隙のところにあったはずの音よりも弱いときに，連続聴錯覚は生じないということである。

空隙転移錯覚 The gap transfer illusion

この錯聴においては，物理的に連続である音が不連続であるように聴こえる (Nakajima et al., 2000; Nakajima, 2006; Kanafuka et al., 2007; Kuroda, Nakajima, & Eguchi, 2008; Remijn et al., 2008)。約 1.5 s 以上の周波数変化音の中央付近に 0.1～0.3 s 程度の時間的空隙を設け，その中央において，0.5 s 程度の周波数変化音と交差させると，この短い音は物理的に連続であるにもかかわらず，空隙を有するように知覚され，物理的に空隙を有する長い音は連続しているように知覚される (Figure 3)。この錯覚は大変鮮やかであり，条件によっては，短い音が物理的に途切れている場合 (Figure 4) と，錯覚によって途切れているように聴こえる場合とを，ちょっと聴いただけでは聴きわけられない。

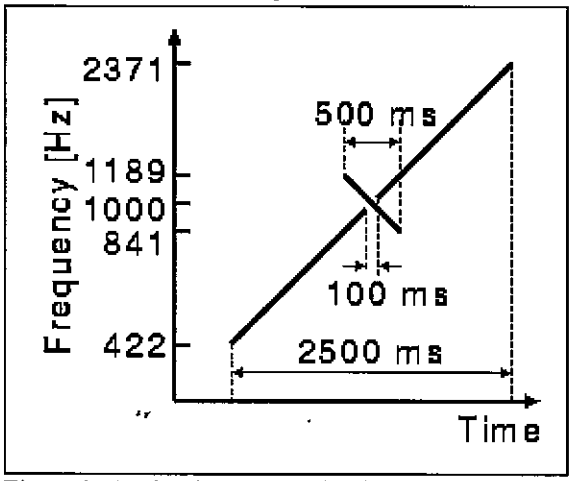


Figure 3. A stimulus pattern that is likely to cause the gap transfer illusion. The gap in the longer tone is perceived as if it were in the shorter tone.

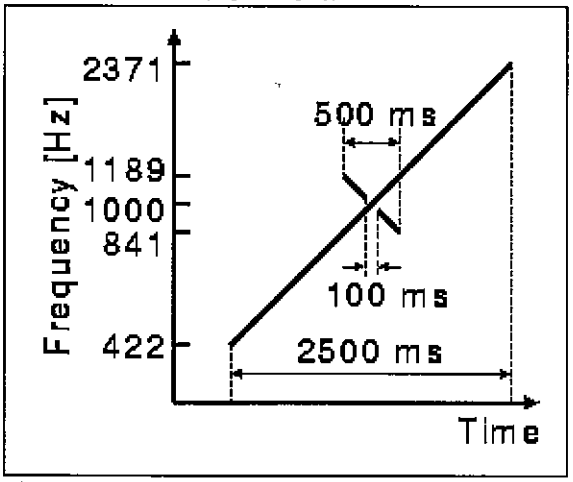


Figure 4. A stimulus pattern to be compared with the stimulus pattern in Figure 3. These two patterns give almost identical percepts.

長い音が物理的に不連続であるにもかかわらず連続であるように知覚されるという面に関して，この錯聴は連続聴錯覚を含んでいるようにも見える。しかし空隙転移錯覚は，空隙を埋める短い音が長い音よりも数デシベル弱い場合にも生ずることが判っており，これが典型的な連続聴錯覚と同じ仕組みによるものとは考えがたい (Kuroda et al., 2008)。

合成音声を用いて，日本語母音の /a/ を周波数変化音とすると，/k/，/w/ などの子音が，空隙と同じように知覚のうえで長い音から短い音に転移する (Tsunashima & Nakajima, 2002; Nakajima, 2008)。

この錯聴は，空隙が長い音と短い音とのいずれに属するかについて，明確な手掛かりが聴覚システムに与えられていないことによって生じているのであって，錯覚と呼ぶ必要はないとの見解もありうる。その見解に対しては，現象にどのような名称を付けるかは本質的な問題ではなく，空隙がどちらの音に知覚される可能性もありながら，なぜ殆どいつも，物理的には空隙を含んでいない短い音が空隙を有するように聴こえるのか，という点が重要であると答えておきたい。どのように名付けようと，説明を要する現象がそこにあることに変わりはない。

空隙の単一帰属化 The illusory gap unification

この錯聴が典型的に生ずる場合，はっきりと検出されうるくらいに長い空隙が複数の音に同時に現れるにもかかわらず，音の一つが連続しているように知覚される (Remijn, Nakajima, & Tanaka, 2007; Remijn et al., 2008)。約 1.5 s 以上の周波数変化音と，0.5 s 程度の反対方向に周波数変化する音とがそれぞれの中央で交差し，交差点において 0.05 s 以下の時間的空隙を共有するとき，この空隙が短い音にのみ知覚され，長い音は連続しているように知覚されることが多い (Figure 5)。長い音に空隙が知覚される場合にも，短い音の空隙に比べて不明瞭になる。

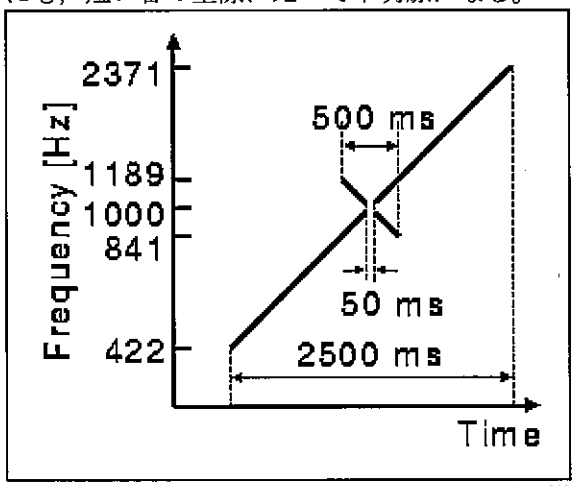


Figure 5. A stimulus pattern that can cause the illusory gap unification. The gap is perceived clearly only in the shorter tone. The continuity of the longer tone will be clearer if the gap is shortened, but sometimes at the cost of the perceptual clarity of the gap in the short tone.

同じように長さの異なる周波数変化音を，物理的に連続したまま互いに逆位相で交差させると，類似の錯聴を認めることができる。この場合，唸りによって交差点の付近に音の強さの落ちこみが生じ，空隙の役割を果たしているが，空隙によるスペクトルの広がりが全く生じない条件においても同様の現象が現れることを示す点で重要である (Nakajima et al., 2000)。

空隙の単一帰属化については，研究を発展させてゆくさまざまな可能性がある。上記の連続聴錯覚の例で用いた雑音の中央にごく短い空隙を挿入しても，依然として純音が連続したものとして知覚されうることを，Ciocca (2007) が示している。定常的な音の高さを有する三つのリコーダー音を合成し，その一つを長い音 (2 s) とし，二つを短い和音 (0.4 s) として，全ての音の中央の時刻を揃え，ここに数十ミリ秒程度 (完全な空白を 0.02 s とする) のごく短い空隙を挿入することができる (Figure 6)。この空隙は三つの音の全てが共有するものである。この際，条件を調整することによって，短い和音のみが空隙を含み，長い音が連続するような知覚を生ずることができる (Nakajima, 2007) 。これまで，主として単一成分の単純な音を用いてきた限りでは，定常的な音の高さを有する音を用いてこの錯聴を起こすことは難しかったので，このデモンストレーションから，新たに研究を展開させることができよう。

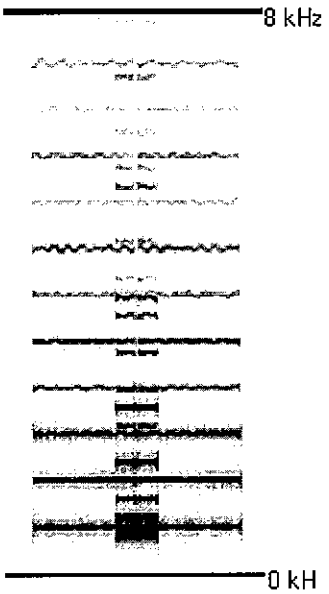


Figure 6. The spectrogram of a stimulus pattern consisting of three synthesized recorder tones. The total duration of the pattern is 2 s. The gap of 0.02 s in the middle is shared by all the components. The percept of the 2-s long tone is clearly continuous.

この錯聴において長い音は，その時間的空隙が何か別の音によって埋められているということがないにもかかわらず，連続しているように聴こえる。この際，単に空隙が短すぎるのではっきりと知覚されないというわけではない。もしそうであるならば，

短い音の空隙もはっきりと知覚されないはずである。ここで長い音が連続しているものと錯覚されていることを，聴覚システムの時間分解能の限界によって説明することは不可能である。

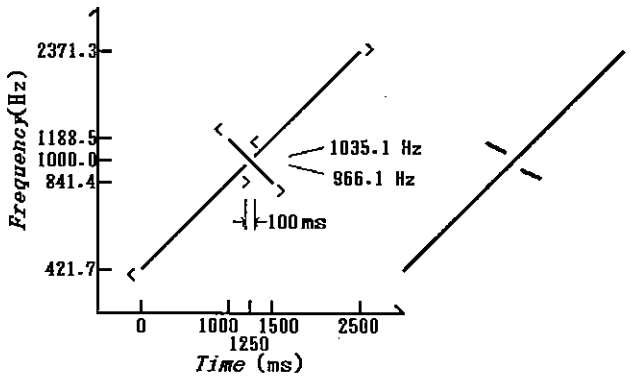


Figure 7. The event construction model. An onset (<) and an offset (>) as indicated in the stimulus pattern in the left panel are connected to each other perceptually if they are close to each other, and if they are presented in this order (<>). Then, a percept as in the right panel appears, which is the gap transfer illusion.

3. 音事象構築モデル

ここで紹介した錯聴を統一的に理解するために，音事象構築モデル the event construction model を導入することができる (Nakajima et al., 2000)。最初に空隙転移錯覚に即してモデルの説明を行う。Steiger の非公式の観察 (Bregman, 1990) は，周波数変化音の始まりと終わりが，知覚のうえで独立した要素であるかのようにふるまいうることを示唆している。筆者もこれを確かめるために，1000 Hz から 2000 Hz まで 1 s かけて上昇するような単一成分の周波数変化音を作り，0.5 s の間隔を音と音とのあいだに挟んで反復呈示した。立上がり時間，立下がり時間は 0.02 s とした。一般に認められているように「音の始まり」に対応する独立した短い音をはっきりと聴きとることができた。しかし，「音の終わり」に対応する音を聴きとることは難しかった 。周波数変化の方向を逆（下降）にしてもほぼ同様の観察結果となった。おそらく「音の終わり」は，それだけで完結した知覚内容になることはありえず，独立に検出された場合には「音の始まり」，「継続部」が知覚のうえで補完されなければ，完結した知覚内容を構成しないのではないかと筆者らは考えている(寺田，中島，2008，参照) 。「音の始まり」，「音の終わり」がそれぞれどのように知覚内容に反映されるかについては，さらに検討が必要であるが，これらが分離しうる知覚要素であるとの仮説を立てること自体は，それほど不自然ではなかろう。さらに，「音の始まり」と「音の終わり」とが知覚のうえで結びつくときには，ある種の文法に従って必ずこの順で結びつくと考えることにする。これは，「音は必ず『音の始まり』から始まる」という当然のことを，仮説として改めて言い表しただけのことである。そして，ゲシタルト心理学で言うところの「近接の原理」に従い，「音の始まり」，「音の終わり」の手掛かりが，時間－周波数の座標において近い場合ほど知覚のうえで結びつきやすいと考える (Nakajima & Sasaki, 1996)。大雑把な仮説ではあるが，これによっていくつかの現象をまとめて理解することが可能になり，さらに精密な心理実験を行うきっかけが得られる。

空隙転移錯覚は，次のように説明することができる (Figure 7)。上記の刺激パターンにおいて，短い周波数変化音の始まりには「音の始まり」の手掛かりが与えられ，長い周波数変化音に挿入された空隙の直前には「音の終わり」の手掛かりが与えられている。これらの手掛かりは，時間，周波数ともに近接しており，知覚のうえで結びついて音を示しうる順序に並んでいる。したがって，この二つの手掛かりが結びついて，一つの音の知覚内容を構成する。長い周波数変化音の空隙の直後には「音の始まり」の手掛かりが与えられ，短い周波数変化音の終わりには「音の終わり」の手掛かりが与えられているので，この二つの手掛かりも同様に結びついて，もう一つの音の知覚内容を構成する。このようにして，短い音が物理的には連続しているにもかかわらず，対応する時間的な位置には二つの短い音が知覚される。言いかえれば，短い音が空隙を含んでいるように聴こえることが説明される。空隙の前後の「音の終わり」，「音の始まり」の手掛かりは知覚のうえで解釈されたので，改めて解釈される必要はなくなり，長い周波数変化音は物理的に不連続であるにもかかわらず，連続しているように知覚される。このようにして，空隙が長い音から短い音に乗りうつったかのように知覚される錯聴が説明された。

この説明は定性的なものであり，今後知覚実験によって定量的な予測ができるようにしてゆく必要がある。しかし，モデルは現象を大まかに把握するうえで有効であり，筆者らはこのモデルの基本となる考えかたに沿って分離音現象 the illusory split-off と名づける別の錯聴を発見した (Nakajima & Sasaki, 1996; Nakajima et al., 2000; Figure 8)。この現象は，二つの周波数変化音が，0.1-0.3 s 程度の時間的な重なりを有するように継時的に呈示されると，重なりに対応すると思われる時間帯に，物理的には存在しない短い音が知覚されるという現象である。加えて，ここで紹介した空隙の単一帰属化も，モデルを考える過程において予測し，発見したものである (Remijn, Nakajima, & Tanaka, 2006)。少なくとも二つの現象を発見することに繋がったのであるから，一時的にはモデルの存在価値があったと言ってよいであろう。

空隙の単一帰属化に関してこれまで明確に述べたことはないが，モデルを適用することは難しいことではない。上記 (Figure 5) の刺激パターンにおいて，空隙の直前，直後では二つの周波数変化音の周波数が極めて近く，その違いは，聴覚システムの末梢に

おける周波数分析の単位である臨界帯域幅よりも充分に小さい。周波数変化音が調波的な複数の成分から成る場合であっても，交差する成分どうしを見れば，この点は同様である。そこで，「音の終わり」，「音の始まり」の手掛かりが，それぞれ一つずつ与えられたと考え，一つの手掛かりは一度だけ解釈されればよいと考えれば，「音の始まり」，「音の終わり」の手掛かりの配置は空隙転移錯覚の場合と類似したものになり，空隙転移錯覚の知覚内容に似た知覚内容の生ずることを説明することができる。ただ問題になるのは，音エネルギーの全くない数十ミリ秒程度の区間においても，長い周波数変化音が連続して，少なくとも明瞭に不連続ではないものとして知覚されるのはなぜかということであろう。ここで注目したいのは，Efron (1970a; 1970b) が，スペクトルの異なる定常音が 0.02-0.04 s の空隙を介して呈示されるとき，それらが明瞭に隔てられたものであると知覚され，音のない区間という意味での空隙が知覚されない場合があることを報告していることである。Nakajima & Sasaki (1996) が，音のない区間の知覚内容を「空白部 silence」と呼ぶことを提案しているのに従わせていただくと，空隙の単一帰属化を生ずるような条件では，「音の始まり」，「音の終わり」の手掛かりのみが与えられ「空白部」の手掛かりが与えられなかったので，長い音の知覚内容に空白部を含める必要がないと考えることができる。もし時間的空隙が 0.04 s よりも充分に長くなれば，たとえば 0.2 s になれば，短い音と長い音との両方に「空白部」があるという手掛かりが与えられるので，空隙の単一帰属化は生じないと考えられる。これは事実に一致する (Remijn, Nakajima, & Tanaka, 2007)。

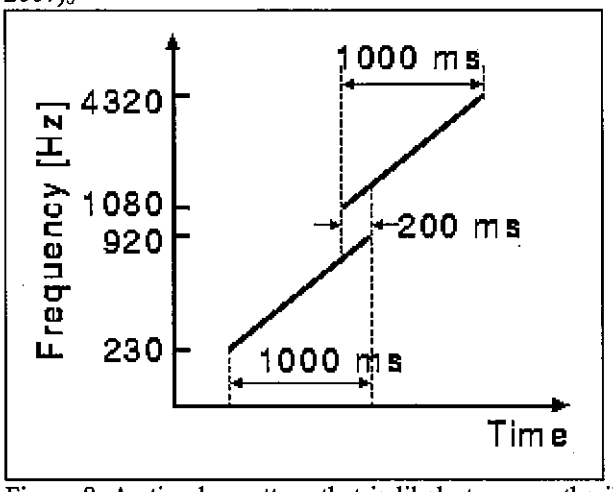


Figure 8. A stimulus pattern that is likely to cause the illusory split-off. Typically, the listener perceives a short tone, probably corresponding to the overlap, in the middle of the pattern. The rest of the percept is often a long glide corresponding to the whole duration of the stimulus pattern.

先に述べたように，広く認められている考えかたの一つによるならば，長い音の一部が空隙となり，そこに短い音がはめこまれて連続聴錯覚が生ずるためには，短い音が，長い音が仮に続いていてもそれをマスキングによってかき消してしまうくらいに強いことが要件である。ところが，空隙転移錯覚や空隙の単一帰属化において，長い音が錯覚として連続しているように知覚されることに関して，この要件は全く満たされていない。いずれの場合にも，空隙が確かに検出されていながら，それがどちらの音に帰属するかに関して錯覚が生じており，聴覚の末梢にその原因を求めることは不可能である。高次のレベルが関わるような仕組みを考えることが必要であり，本稿に取りあげたモデルは，大雑把なものではあるが，考えを進める出発点になるであろう。

よく知られている連続聴錯覚についても，考えを示しておきたい (Nakajima & Sasaki, 1996; Remijn & Nakajima, 2005)。具体的には，音事象構築モデルを連続聴錯覚に適用することが可能であるかどうかを検討する。純音の一部がごく短い空隙となり，純音の周波数を含む充分に強い雑音がこの空隙を埋める場合 (Figure 1)，空隙の前後に生ずる純音の「音の終わり」，「音の始まり」の手掛かりはマスキングによってかき消されるので，ここには挿入された雑音の「音の始まり」，「音の終わり」の手掛かりのみが与えられている。この二つの手掛かりは，時間，周波数のうえで近接しているので，知覚のうえで結びついて雑音の知覚を生ずるはずである。そうするとこれらの手掛かりは改めて解釈されることを要せず，また，この付近には「空白部」の手掛かりが与えられてはいないので，純音の中に空隙を知覚しなければならない何の理由もなく，純音は連続したものとして知覚されるはずである。このように，同じモデルによって連続聴錯覚を説明することができる。

連続聴錯覚に関しては多くの研究がなされており，その生ずる仕組みについても様々なことが言われている。特に，短い音に対して生じた末梢の興奮の一部が，長い音の空隙を補完するために使われ，短い音の音エネルギーの一部が長い音の一部として知覚されるように見える，との考えかたは解りやすく有益である (例えば Warren, 2008)。一方，ここで紹介したモデルはいくつかの錯覚現象を統一的に説明する点で優れている。連続聴錯覚に関しては，一方で，同じ仕組みを二つの異なったモデルによって捉えている可能性があり，他方で，異なる二つ以上の仕組みが働いている可能性もある。本稿においては，音事象構築モデルが単純である割には多くを説明しうるものであることを指摘しておきたい。もちろん，定量的なデータが決定的に不足していることは問題であるが，これからどのような実験を行えばよいかという指針として捉えれば，モデルとしての役割は果たしうるであろう。

4. おわりに

音事象構築モデルで表されるような特性を，なぜ

ヒトの聴覚が持っているのかを考えることによって，ヒトが与えられた手掛かりをどのようにまとめあげて，物音，話し言葉，音楽などをつながりのある音として聴き，場合によっては他の音との関係を把握するのかについて，理解が進むであろう。ロボットに聴覚を組みこむという目的に沿うべく話題を準備したが，このあとどこから手をつけてよいのかが解りにくいかもしれない。素人の意見ではあるが二つのことを述べさせていただきたい。一つは，ロボットに組みこむ聴覚システムがヒトの聴覚システムを真似る必要はないということである。脳とコンピューターとのあいだには似た面がまず目につくかもしれないが，異なる点も多い。ロボットの中枢がコンピューターであるならば，それに適した聴覚システムがあってよいように思う。もう一つは，矛盾するようであるが，ヒトの聴覚システムを丹念に模倣してゆき，ロボットという全体の中で動かすことによって，ヒトの聴覚についての理解が深まるであろうということである。定量的なデータが欠けていても適宜推測して全体を作りあげることによって，今後心理学，生理学の分野においてどのような実験を優先して行うべきかの見当を付けやすくなるのではなかろうか。

ここで紹介した話題の多くは，聴覚コミュニケーションについて考察する際に避けては通ることができないと思われる (Tsunashima & Nakajima, 2002; Nakajima, 2008) 。言語を聴きとる際に，どこからどこまで音がつながっていて，どこで切れているのかが判らなければ，文字通り話にならないであろう。音楽においては，音と音の時間的なつながりがメロディーになるので，まず一つひとつの音が知覚のうえでどのように決定されるかを理解するために，本稿で紹介したようなモデルを拡張したものが必要になるであろう。もし未来に人間とロボットとが一つの社会を作るのであれば，これから人間の聴覚を作りかえるわけにゆかない以上，ロボットにも人間の聴覚コミュニケーションに参加することができるような聴覚を持ってもらう必要がある。その聴覚は，同時に危険や変化を察知するという，全く異なった役割をも果たさなければならない。ロボットにも真似のしやすい聴覚の基本原理を捉えることが重要であり，あるいは音事象構築モデルが役立つかもしれない。

謝辞）本発表の理論的な枠組みは，佐々木隆之氏との共同研究によって作りあげたものである。また，上田和夫氏との討論による部分も多い。科学研究費補助金の援助を受けた (平成 19-23 年度 19103003，平成 20-22 年度 20653054)。プレゼンテーションの作成には，廣瀬有希子氏の協力を得た。

註）本稿は，今年の７月から８月にかけてトロント（カナダ）で開催された The 24th Annual Meeting of the International Society for Psychophysics において筆者が行った講演の予稿に，図を加えるなど内容を追加したものである。

参考文献

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

Ciocca, V. (2007). Personal communication. December.

Ciocca, V. & Bregman, A. S. (1987). Perceived continuity of gliding and steady-state tones through interrupting noise. *Perception & Psychophysics*, 42, 476-484.

Efron, R. (1970a). The relationship hetween the duration of a stimulus a stimulus and the duration of a perception. *Neuropsychologia*, 8, 37-55.

Efron, R. (1970b). The minimum duration of a perception. *Neuropsychologia*, 8, 57-63.

Kanafuka, K., Nakajima, Y., Remijn, G. B., Sasaki, T., & Tanaka, S (2007). Subjectively divided tone components in the gap transfer illusion. *Perception & Psychophysics*, 69, 641-653.

Kuroda, T., Nakajima, Y., & Eguchi, S. (2008). Effects of the sound-pressure-level difference between crossing glides on the occurrence of the gap transfer illusion. *Proceedings of the 24th Annual Meeting of the International Society for Psychophysics, Toronto, Canada*. 53-58.

Miller, G. A. & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *Journal of the Acoustical Society of America*, 22, 167-173.

Nakajima, Y. (2006). Demonstrations of the gap transfer illusion. *Acoustical Science and Technology*, 27, 322-324.

Nakajima, Y. (2008). Illusions related to auditory grammar: Ten demonstrations in musical contexts. *Proceedings of the 10th International Conference on Music Perception and Cognition, Sapporo, Japan*, 301-304.

Nakajima, Y., & Sasaki, T. (1996). A simple grammar of auditory stream formation [Abstract]. *Journal of the Acoustical Society of America*, 100, 2681.

Nakajima, Y., Sasaki, T., Kanafuka, K., Miyamoto, A., Remijn, G., & ten Hoopen, G. (2000). Illusory recouplings of onsets and terminations of glide tone components. *Perception & Psychophysics*, 62, 1413-1425.

Remijn, G. B., & Nakajima, Y. (2005). The perceptual integration of auditory stimulus edges: An illusory short tone in stimulus patterns consisting of two partly overlapping glides. *Journal of Experimental Psychology: Human Perception & Performance*, 31, 183-192.

Remijn, G., Nakajima, Y., & Tanaka, S. (2007). Perceptual completion of a sound with a short silent gap. *Perception*, 36, 898-917.

Remijn, G. B., Pérez, E., Nakajima, Y., & Ito, H. (2008). Frequency modulation facilitates (modal) auditory restoration of a gap. *Hearing Research*, 243, 113-120.

Riecke, R., van Opstal, A. J., & Formisano, E. (2008). The auditory continuity illusion: A parametric investigation and filter model. *Perception & Psychophysics*, 70, 1-12.

寺田知未, 中島祥好 (2008). 音の終わりから錯覚的に音事象が構築される可能性について. *日本音響学会聴覚研究会資料* (印刷中).

Tsunashima, S., & Nakajima, Y. (2002). Demonstrations of the gap transfer illusion. *Proceedings of the 7th International Conference on Music Perception and Cognition, Sydney, Australia*, 407-410.

Warren, R. M. (2008). *Auditory Perception: An Analysis and Synthesis: Third Edition*. Cambridge: Cambridge University Press.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A802-2 (11/18)

Vocal Imitation Model with Segmenting and Composing Capability of Vowel Structure using Recurrent Neural Network

Hisashi Kanda, Tetsuya Ogata, Toru Takahashi, Kazunori Komatani and Hiroshi G. Okuno
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University
Engineering Building #10, Sakyo, Kyoto, 606-8501, JAPAN
{hkanda, ogata, tall, komatani, okuno}@kuis.kyoto-u.ac.jp

Abstract

This paper shows a continuous vocal imitation system using a computational model that explains the process of phoneme acquisition by infants. Human infants perceive speech sounds as continuous acoustic signals and segments them into phonemes with articulatory movement. To repulicate the development process, we apply the segmenting method to our system by Recurrent Neural Network with Parametric Bias (RNNPB). This method determines the multiple segmentation boundaries in a sequence using the prediction error of RNNPB, and the PB values of RNNPB can be encoded as kind of phonemes. The method is implemented into a physical vocal tract model, called Maeda model. Experimental results demonstrated that our system can imitate vocal sound involving arbitrary numbers of vowels using the vowel structure in the RNNPB.

1 Introduction

Our goal is to clarify how to acquire the ability to distinguish phonemes in early human infants. Human infants can acquire spoken language through vocal imitation of their parents. Despite their immature bodies, they can imitate their parents' speech sounds by generating those sounds repeatedly. This ability is closely related to the cognitive development of language.

Many researchers have designed vocal imitation systems that duplicate the developmental process of infants' vowel acquisition [1] [2] [3]. These studies were based on the idea that articulatory mechanisms such as the vocal tract enable us to acquire phonemes. This idea has been advocated as the *motor theory of speech perception* [4], and recent neuroscience studies seem to show the idea to be an active process involving motor cognition [5] [6].

Segmenting acoustic signals with articulatory movements is essential for vocal imitation and phoneme acquisition; the reason is that human infants do not know the given phonetic distinction inherently. The human development studies described above assume that acoustic signals consist of discrete phoneme sequences in advance, and they search for vocal tract shapes corresponding to phonemes. However, articulatory movements for the same phoneme dynamically change according to the context of continuous speech (e.g. coarticulation). This effect derives from a physical constraint that articulatory movements should be continuous in sound generation. We assume that human infants regard phoneme sequences as continuous acoustic signals. As they grow, infants will acquire the ability to discover phoneme units in a continuous speech sound by prosody, rhythm, stress and whether they can imitate the sound or not.

We use Recurrent Neural Network with Parametric Bias (RNNPB) [7] to segment and imitate a continuous temporal sequence consisting of acoustic signal with articulatory movement. From the view point of considering sounds as temporal sequences, we have already developed a vocal imitation system [8], which used the RNNPB model and a physical vocal tract model, called Maeda model, to simulate the physical constraints. We, furthermore, apply to our system the segmenting method by RNNPB [9]. This method can segment several kinds of sequences into primitive sections using the prediction error of the RNNPB model and encode the segmented sections as a set of parameters, called PB values. It is assumed that the method enables our system to encode the position of phoneme transition as the segmented sections, and that imitating heard sounds, our imitation system can manipulate the encoded phonemes.

2 Vocal Imitation Process and Model

2.1 Overview of Our Imitation Process

As illustrated in Fig. 1, our imitation process consists of three phases: learning, recognition, and generation.

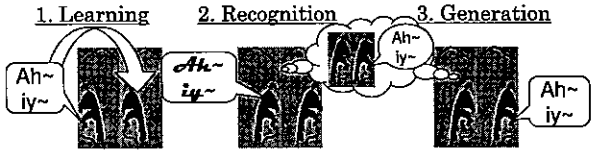


Figure 1: Imitation process.

1. Learning (Babbling)
   Our system uses articulatory movements to produce sounds, and it makes a connection between the movement and the produced sound. This phase corresponds to babbling in infants.
2. Recognition (Hearing parents' speech sounds)
   We put speech sounds into the system. The system recognizes the sounds with articulatory movements producing the same dynamics as the input sound.
3. Generation (Vocally imitating heard sounds)
   Finally, the system uses the articulatory movement to imitate a speech sound.

The learning phase uses the RNNPB method of segmenting temporal sequences. Our imitation model can self-organize so as to connect an articulatory movement with the corresponding sound dynamics. Additionally, in the recognition and generation phases, the connection is available for our model to imitate speech sounds.

2.2 Physical Vocal Tract Model

A synthesizer simulating human vocal tract incorporates the physical constraints of the articulatory mechanism. The vocal tract parameters with physical constraints are better for continuous speech synthesis than acoustic parameters such as the sound spectrum. This is because the temporal change of the vocal tract parameters is continuous and smooth, while that of the acoustic parameters is complex, and it is difficult to interpolate the latter parameters between phonemes.

We used the vocal tract model proposed by Maeda [10]. This model has seven vocal tract parameters: 1. Jaw position (JP), 2. Tongue dorsal position (TDP), 3. Tongue dorsal shape (TDS), 4. Tongue tip position (TTP), 5. Lip opening (LO), 6. Lip protrusion (LPR), 7. Larynx position (LP). The parameters were derived by principal components analysis of cineradiographic and labiofilm data from French speakers. Although there are other vocoders, such as PARCOR [11] and STRAIGHT [12], we think that Maeda model, with its physical constraints based on anatomical findings, is the most appropriate, because of our aim to simulate the development process of infant's speech. This model for generating acoustic signals is a very simplified articulatory model, and the sound units corresponding to phonemes are expressed in these articulatory terms.

Table 1 shows the first and second formant (F1, F2) of vowels produced by Maeda model. Each Maeda parameter takes on a real value between -3 and 3 and may be regarded as a coefficient weighting an eigenvector. The sum of these weighted eigenvectors is a vector of points in the midsagittal plane that defines the outline of the vocal tract shape. The resulting vocal tract shape is transformed into an area function, which is then processed to obtain the acoustic output and spectral properties of the vocal tract during speech.

Table 1: The F1 and F2 averages of Maeda model .

	/a/	/i/	/u/	/e/	/o/
F1	667	234	269	401	500
F2	1214	2161	924	1894	902

2.3 Learning Algorithm

This subsection describes the method to learn and segment temporal sequence dynamics. We apply the RNNPB model, which was first proposed by Tani [7] as the forwarding forward model. It generates complex movement sequences, which are encoded as the limit-cycling dynamics and/or the fixed-point dynamics of the RNN.

2.3.1 RNNPB model

The RNNPB model has the same architecture as the conventional Jordan-type RNN model [13], except for the PB nodes in the input layer. Unlike the other input nodes, these PB nodes take a constant value throughout each temporal sequence and are used to implement a mapping between fixed-length values and temporal sequences. Figure 2 shows the network configuration of the RNNPB model.

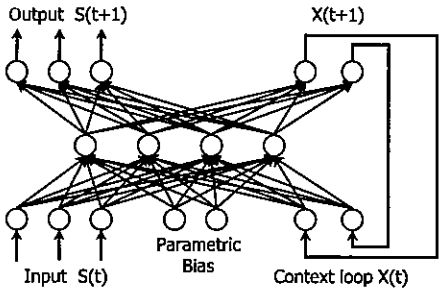


Figure 2: RNNPB model.

Unlike the Jordan-type RNN model, the RNNPB self-organizes the values in the PB nodes that encode the sequence during the learning process. The common structural properties of the training data sequences are acquired as connection weights by using the back propagation through time (BPTT) algorithm [14], as in a conventional RNN. Meanwhile, the specific properties of each individual temporal sequence are simultaneously encoded as PB values. As a result, the RNNPB model self-organizes a mapping between the PB values and the temporal sequences.

2.3.2 Segmenting Temporal Sequence Data

Our segmenting method determines the segmentation boundaries using the prediction error of the RNNPB model. Systems using this approach usually consist of dynamic recognizers that predict the target sequences. The dynamic sequence is articulated based on the predictability of the recognizer. The method we used to segment acoustic signals with articulatory movements uses the prediction error of RNNPB model and the number of segmentations. Its description is as follows: Consider the problem of segmenting a dynamic sequence, $D(t)$, whose length is T into N sections, which are represented as S_i $(i = 0, \cdots, N-1)$. The boundary step between S_{i-1} and S_i is represented by $t = s_i$, that is, S_i is defined as $[s_i, s_{i+1}]$. The segmenting process consists of five steps.

Step 1: Initialization

The given sequence is divided into N sections. Each section has the same length. The boundary step s_i $(i = 0, \cdots, N)$ is set as follows.

$$s_i \longleftarrow i \cdot T/N \tag{1}$$

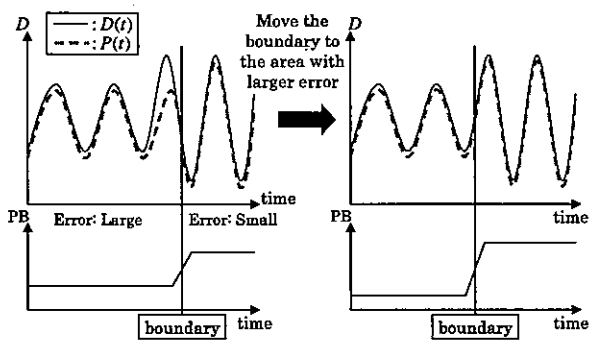


Figure 3: Segmenting multiple dynamics.

Step 2: RNNPB training
The connection weights and PB values of the RNNPB model are updated with the given sequence, while the PB values are kept constant in each section, S_i.

Step 3: Calculate prediction errors
In each S_i, the prediction sequences of the RNNPB model, $P(t)$, are calculated, and the average prediction errors, E_i, is obtained as follows.

$$E_i \longleftarrow \frac{1}{s_{i+1} - s_i} \cdot \sum_{t \in S_i} ||D(t) - P(t)|| \qquad (2)$$

Step 4: Update the length of each section
The boundary step s_i $(i = 1, \cdots, N-1)$ is updated by using the following rules:

$$s_i \longleftarrow \begin{cases} s_i - ds & if\ E_{i-1} \geq E_i \\ s_i + ds & if\ E_{i-1} \leq E_i, \end{cases} \qquad (3)$$

where ds is a parameter used to update the section length.

Step 5: Repeat Steps 2 to 4 until the whole error is less than the threshold.

If a sequence is generated by using simple dynamics, the prediction error of the RNNPB will be small, even when the PB values are fixed. However, if a sequence is generated by using multiple dynamics, the prediction error at the boundary between dynamics will increase as shown in Fig. 3. The algorithm can decrease the error by modifying the position of each boundary.

2.3.3 Learning of PB Vectors

The learning algorithm for the PB vectors is a variant of the BPTT algorithm. The step length of ith section S_i in a sequence is denoted by $s_{i+1} - s_i$. For each of the articulatory and sound parameters outputs, the back-propagated errors with respect to the PB nodes are accumulated and used to update the PB values. The update equations for the kth unit of the parametric bias at the section S_i in the sequence are as follows:

$$\delta\rho_{i,k} = \varepsilon \cdot \sum_{t=s_i}^{s_{i+1}} \delta_{i,k}(t), \qquad (4)$$

$$p_{i,k} = sigmoid(\rho_{i,k} + \delta\rho_{i,k}), \qquad (5)$$

where ε is a coefficient. In Eq. 4, the δ force for updating the internal values of the PB $\rho_{i,k}$ is obtained from the sum of the delta errors $\delta_{i,k}$. The delta error $\delta_{i,k}$ is backpropagated from the output nodes to the PB nodes:

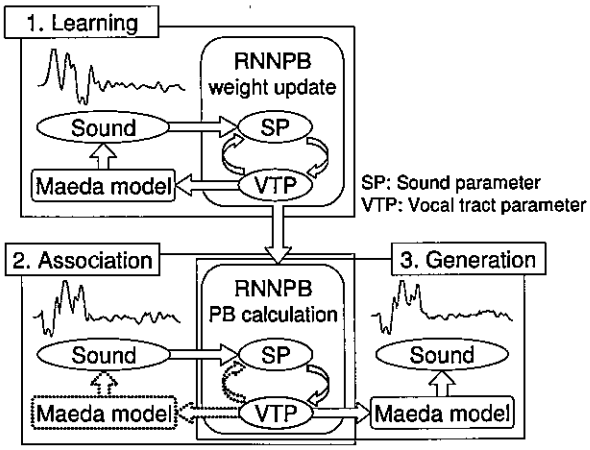


Figure 4: Diagram of the experimental system.

it is integrated over the period from s_i to s_{i+1} steps. Then, the current PB values $p_{i,k}$ are obtained from the sigmoidal outputs of the updated internal values.

2.4 Calculation in Recognition and Generation Phases

After the RNNPB model is organized in the learning phase, it is used in the recognition and generation phases.

The recognition phase corresponds to how infants recognize sounds presented by parents, i.e. to how the PB values are obtained. The PB values of each section are calculated from Eq. 4 and 5 by using the organized RNNPB without updating the connection weights. The boundary steps of each sequence are determined by the prediction errors of the organized RNNPB. However, there is no vocal articulatory data because the system is only hearing sounds without articulating them, unlike in the learning phase. The initial vocal tract values are input to vocal tract units of the input layer in step 0, and the outputs are calculated forward in the closed-loop mode from step 1. More generally, the outputs in the articulatory output layer in step $t-1$ are the input data in the articulatory input layer in step t. This calculation is called *closed loop calculation.*

The generation phase corresponds to what articulation values are calculated. The articulatory output of the RNNPB model is obtained in a *closed loop calculation.* The PB values obtained in the recognition phase are input to the RNNPB in each step.

3 Vocal Imitation System

3.1 Experimental System

Our experimental system is illustrated Fig. 4. This system was used to verify the relation between vocal imitation and the phoneme acquisition process. To simplify the system, we purposely used a simple vocal tract model and target vowel sound segmentation.

In the learning phase, we first use a cubic interpolation method to produce sequences of Maeda parameters as articulatory movements. Second, the sequences are put into Maeda model to produce the corresponding sounds, which are then transformed into temporal sound parameters. Finally, the RNNPB learns each the sound and the Maeda parameters, which are normalized and synchronized. In this phase, the parameter ds was set at 0.1.

Table 2: Input sound data in the recognition phase.

two-vowel	three-vowel	four-vowel
/ae/ /io/	/aeo/ /eai/ /iae/ /oae/ /uai/	/aiue/
/ai/ /iu/	/aeu/ /eia/ /iai/ /oai/ /uao/	/eoai/
/ao/ /oa/	/aia/ /eiu/ /ieo/ /oao/ /uea/	/iueo/
/au/ /oe/	/aie/ /eoa/ /ioa/ /oau/ /uei/	/oaiu/
/ea/ /oi/	/aio/ /eoe/ /ioe/ /oei/ /ueo/	/ueoa/
/ei/ /ou/	/aiu/ /eoi/ /iua/ /oeo/ /ueu/	
/eo/ /ua/	/aoa/ /eou/ /iue/ /oiu/ /uio/	
/eu/ /ue/	/aou/ /eua/ /iui/ /oue/ /uiu/	
/ia/ /ui/	/aue/ /eue/ /iuo/ /oui/ /uoa/	
/ie/ /uo/		

The size of the RNNPB model and the time interval of the sequence data differed according to the experiment.

In the recognition phase, sound data is put into the system. The corresponding PB values are calculated for the given sequence by the organized RNNPB to associate the articulatory movement with the sound data.

In the generation phase, the system generates imitation sounds by inputting the PB values obtained in the recognition phase into the organized RNNPB.

3.2 Sound Parameter

We use a kind of Mel-Frequency Cepstrum Coefficients (MFCCs) as sound parameter, which are obtained from power spectrum of sound waveform. The power spectrum is calculated by STRAIGHT analysis instead of short term Fourier transform of the segment. The power spectrum has no interference caused by fundamental frequency of vocal source. In this paper, MFCC stands for this kind of MFCC. The MFCCs are calculated by taking the Discrete Cosine Transform of mel-scaled log filterbank energies. STRAIGHT analysis is a kind of pitch analysis in which the window length in the analysis is set depending on the fundamental frequency of the sound.

In our experiment, the speech signals were single channel with a sampling frequency 10 kHz. The number of mel filterbank was set to 12. We formed 5-dimensional vectors from low-third to low-seventh dimension out of 12-dimensional MFCC vectors. The vectors produced from speech sounds remain vowel features, and they are almost independent of speakers.

3.3 Vocal Tract Parameter

We applied Maeda model with the first six parameters described in 2.2 section. The reason for choosing only these six parameters is that when Maeda model produces vowel sounds, the seventh parameter LP has a steady value. In the generation phase, it is possible for the Maeda parameters produced by the RNNPB to temporally fluctuate without human physical constraints. This occurs if the system does not easily associate the articulatory movements of an unexperienced sound. Therefore, to help prevent extraordinary articulation, we temporally smoothed the Maeda parameters produced by the RNNPB. Concretely, the Maeda parameters in each step were calculated by averaging those of the adjacent steps.

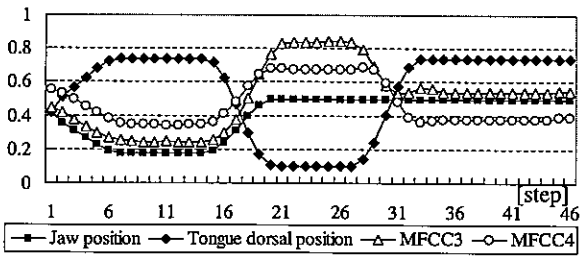


Figure 5: Learning data: /aiu/.

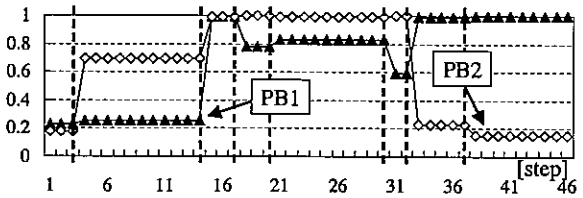


Figure 6: The PB values of /aiu/ in the learning phase.

4 Imitation Experiment

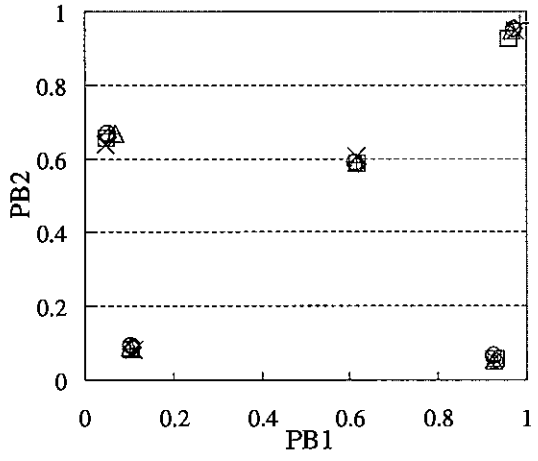
We carried out an experiment of vocal imitation using our system. The organization of RNNPB is as follows: 11 input/output nodes, 40 hidden nodes, 5 context nodes, and 2 PB nodes. In this experiment, we set $ds = 0.1$ and $N = 8$. In the learning phase, RNNPB learned 10 patterns of three-vowel data consisting of the 5-dimensional MFCC vector and the 6-dimensional vocal tract parameters: /aiu/, /aoe/, /iue/, /iao/, /ueo/, /uia/, /eoa/, /eui/, /oai/, and /oeu/ (1350-ms and 30-ms/step), produced by Maeda model.

In the recognition phase, we input the MFCC parameters of the two-vowel, three-vowel and four-vowel sounds, which are produced by 4 people (3 males and 1 female), into the organized RNNPB, and recorded the PB values and the boundary steps for each sound. Table 2 lists input sound data in the recognition phase. The two-vowel data were 900-ms, the three-vowel data were 1350-ms, and the four-vowel data were 2000-ms. We set $N = 4$ in recognizing two-vowel data, and $N = 8$ in recognizing three-vowel and four-vowel data. In the generation phase, we used the PB values and the boundary steps to reproduce the recorded sounds.

Figure 5 shows 4 sequences (JP and TDP of Maeda model, and the third and fourth MFCC) of the learning data /aiu/. Figure 6 shows the PB values for the learning data /aiu/ obtained by the organized RNNPB. The vertical dotted line represents the boundary step s_i segmented by RNNPB in the learning phase. The boundary steps, dividing the input sequence /ueo/ into flat and transition segments, in Fig. 6 were $s_1 = 3$, $s_2 = 14$, $s_3 = 17$, $s_4 = 20$, $s_5 = 30$, $s_6 = 32$ and $s_7 = 37$. We confirmed that as the size of N increases, the boundary steps become more stable in the learning phase. Similar results were also acquired for the other input data.

Figure 7 shows the PB space of the organized RNNPB. In Fig. 7, the PB values represent the phonemes of a set of three-vowel data aligned according to the length of the three longest sections of a learning sequence. The PB values for the same vowel, including the learning data, were mapped with sufficient dispersion.

Figure 8 shows the analysis of PB space. This analysis was conducted as follows:



□ /a/ (/aiu/) ◇ /a/ (/eoa/) △ /a/ (/oai/) × /a/ (/iao/) ○ /a/ (/aoe/) + /a/ (/uia/)
□ /i/ (/aiu/) ◇ /i/ (/iue/) △ /i/ (/oai/) × /i/ (/iao/) ○ /i/ (/eui/) + /i/ (/uia/)
□ /u/ (/aiu/) ◇ /u/ (/iue/) △ /u/ (/ueo/) × /u/ (/oeu/) ○ /u/ (/eui/) + /u/ (/uia/)
□ /e/ (/iue/) ◇ /e/ (/ueo/) △ /e/ (/eoa/) × /e/ (/aoe/) ○ /e/ (/oeu/) + /e/ (/eui/)
□ /o/ (/ueo/) ◇ /o/ (/eoa/) △ /o/ (/oai/) × /o/ (/iao/) ○ /o/ (/aoe/) + /o/ (/oeu/)

Figure 7: The PB space in the learning phase.

1. The PB space was divided into 10 x 10 lattices.
2. For each lattices, each sequence of Maeda parameters was obtained through *closed loop calculation.*
3. Using Maeda parameter sequences, 300-ms speech sounds were produced ($N = 1$).
4. The F1 and F2 averages of second half of each produced sound were calculated.
5. The square error of F1 and F2 averages from those of Table 1 were calculated for each vowel.
6. The vowel corresponding to the minimum square error was set at each lattice point.

In Fig. 8, each color expresses the vowel: blue is /a/, red is /i/, yellow is /u/, green is /e/, and purple is /o/. In the space, bright color represents small error, and dark color represents big error. The PB values corresponding to constant vowels in Fig. 7 are plotted on the bright place in Fig. 8. Each vowel has nonlinear distribution due to F1 and F2 formants. Especially, the vowel /a/ is widely distributed in the PB space.

Figure 9 shows the transition of the PB values for the input data /aeo/ and /eoa/ of one male in the recognition phase. In Fig. 9, the PB values of section S_1 for /aeo/ separated from those of the sections $S_{6,7}$ for /eoa/. We confirmed that the category of the phoneme /a/ in Fig. 8 corresponded to the transitions of the PB values in Fig. 9.

In the generation phase, most of imitation sounds were similar to the original. It is confirmed that the PB values of each vowel obtained in the recognition phase correspond with those in the learning phase. Figure 10 shows the F1-F2 map for each vowel of one male's two-vowel imitated sounds. In Fig. 10, the formants of imitated sounds except for /o/ correspond with those of human speech sound (the map of "vowel triangle" shown in [15]).

It is confirmed that our model can imitate vocal sound involving arbitrary numbers of vowels using the vowel space in the RNNPB. The space is acquired by "babbling" of the vocal tract model with only a few sets of vowel sounds.

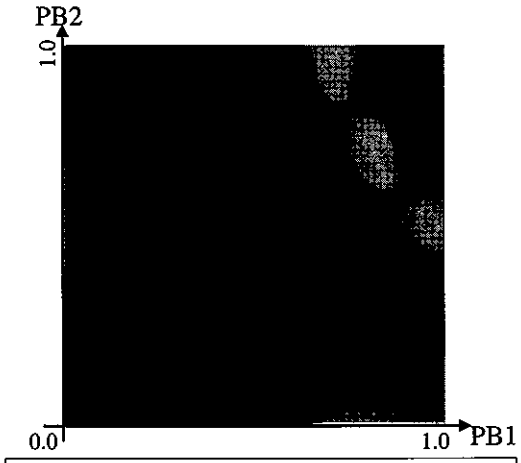


Figure 8: Analysis of PB space.

5 Discussion

5.1 Vowel acquisition

Our system could encode the same vowels in acoustic signals as the near PB values in the PB space. In this sense, each vowel category is defined independently from the other vowels. However, in Fig. 8, it is confirmed that each vowel category is widely distributed. In Fig. 9, the PB sequences pass through different points in the same vowel categories. This means that the PB values representing the same vowel are changed by the adjacent vowels in a given vowel sequences. It is assumed that this represents coarticulation designed in general speech recognition systems. In this sense, each vowel is determined context dependently on the other vowels.

Tani et al. showed that the internal symbolic process was embedded in the dynamical attractor in a mobile robot system [16]. In his experiment, the robot acquired attractors representing the observed objects as activities in RNN nodes. These attractors were also represented by complex clusters, and the positions of active points were fluctuated by the context, i.e. trajectory of mobile robot. This bilateral characteristic, that is context dependency or independency, is one of the interesting and essential properties in dynamical systems representation.

5.2 Vowel imitation

Our system could accurately reproduce, to an extent, most of the heard sounds that were experienced or unexperienced. In the experiment, information of the fundamental frequency (F0) was eliminated from input sound parameters. Due to this elimination, our system could imitate many vowel patterns of heard sounds that were experienced or unexperienced. In Fig. 9, our system could manipulate the PB values as vowels, and robustly recognize the context of sound.

6 Conclusions and Future Works

We developed a vocal imitation system applying the segmenting method based on predictability by RNNPB. Through the experiment, the segmenting method enables our system to self-organize vowel space as the PB space without information of the number and kinds of

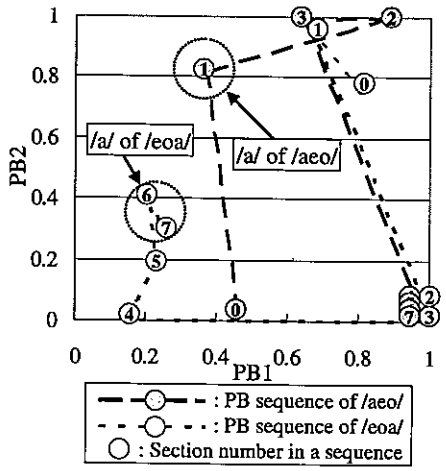


Figure 9: PB sequences for input data */aeo/* and */eoa/* in the PB space.

vowels for input acoustic signals. Furthermore, imitating heard sounds, our system can manipulate the PB values as vowels in the organized PB space. For example, learning only 10 pattern of three-vowel data enables our system to imitate two-vowel and four-vowel sounds in spite of unexperienced vowel sequences. In the analysis of the organized PB space, it is confirmed that each vowel has widely distribution in the PB space and that the distribution expresses the context of speech sounds.

Our future work includes to imitate speech sounds through simulating mother and child interaction. The babbling should be introduced into our model as the exploring and learning phase of corresponding between generated acoustic signal and articulatory movements.

Acknowledgement

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (S), and Creative Scientific Research.

References

[1] B. de Boer, "Self-organization in vowel systems," *Journal of Phonetics*, vol. 28, no. 4.

[2] P. Y. Oudeyer, "The self-organization of speech sounds," *Journal of Theoretical Biology*, vol. 233, no. 3, pp. 435–449, 2005.

[3] K. Miura, M. Asada, K. Hosoda, and Y. Yoshikawa, "Vowel acquisition based on visual and auditory mutual imitation in mother-infant interaction," in *ICDL2006*, 2006.

[4] A. M. Liberman, F. S. Cooper, and et al., "A motor theory of speech perception," in *Proc. Speech Communication Seminar, Paper-D3*, Stockholm, 1962.

[5] L. Fadiga, L. Craighero, G. Buccino, and G. Rizzolatti, "Speech listening specifically modulates the excitability of tongue muscles: a tms study," *European Journal of Cognitive Neuroscience*, vol. 15, pp. 399–402, 2002.

[6] G. Hickok, B. Buchsbaum, C. Humphries, and T. Muftuler, "Auditory-motor interaction revealed by fmri," *Area Spt. Journal of Cognitive Neuroscience*, vol. 15, no. 5, pp. 673–682, 2003.

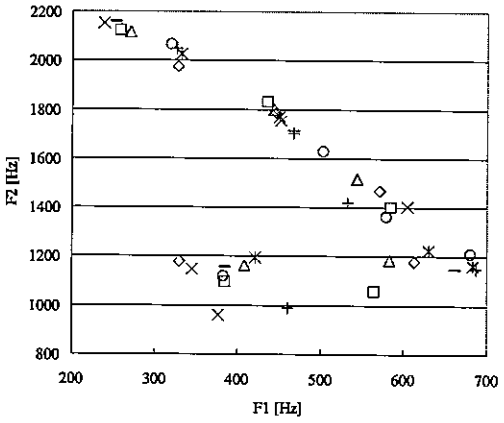


Figure 10: The F1-F2 space in the recognition phase.

[7] J. Tani and M. Ito, "Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment," *IEEE Transactions on SMC Part A*, vol. 33, no. 4, pp. 481–488, 2003.

[8] H. Kanda, T. Ogata, K. Komatani, and H. G. Okuno, "Vocal imitation using physical vocal tract model," in *IEEE/RSJ IROS2007*, 2007.

[9] T. Ogata, M. Murase, J. Tani, K. Komatani, and H. G. Okuno, "Two-way translation of compound sentences and arm motions by recurrent neural networks," in *IEEE/RSJ IROS2007*, 2007.

[10] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," pp. 131–149, 1990.

[11] N. Kitawaki, F. Itakura, and S. Saito, "Optimum coding of transmission parameters in parcor speech analysis synthesis system," *Transactions of the Institute of Electronics and Communication Engineers of Japan (IEICE)*, vol. J61-A, no. 2, pp. 119–126, 1978.

[12] H. Kawahara, K. Masuda, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction," vol. 27, pp. 187–207, 1999.

[13] M. Jordan, "Attractor dynamics and parallelism in a connectionist sequential machine," pp. 513–546, 1986.

[14] D. Rumelhart, G. Hinton, and R. Williams, *Learning internal representation by error propagation.* Cambridge, MA, USA: MIT Press, 1986.

[15] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals.* Prentice-hall, 1978.

[16] J. Tani, "Model-based learning for mobile robot navigation from the dynamical systems perspective," *IEEE Trans. on Systems, Man, and Cybernetics Part B: Cybernetics*, vol. 26, no. 3, pp. 421–436, 1996.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A802-3 (11/18)

ビートトラッキングロボットの構築と評価
A Beat-Tracking Robot and Its Evaluation

村田 和真 †1, 中臺 一博 †1,2, 武田 龍 †3, 奥乃 博 †3, 長谷川 雄二 †2, 辻野 広司 †2
Kazumasa MURATA†1, Kazuhiro NAKADAI,†1,2 Ryu TAKEDA†3, Hiroshi G. OKUNO†3,
Yuji HASEGAWA†2, and Hiroshi TSUJINO†2
†1 東京工業大学, †2(株) ホンダ・リサーチ・インスティチュート・ジャパン, †3 京都大学
†1Tokyo Institute of Technology, †2Honda Research Institute Japan Co., Ltd., †3Kyoto University
murata@cyb.mei.titech.ac.jp, nakadai@jp.honda-ri.com

Abstract

Musical beat tracking is one of the effective technologies for human-robot interaction such as musical sessions. This paper addresses a musical beat tracking robot which can step, scat and sing according to musical beats by using its own microphone. To realize such a robot, we propose a beat tracking method by introducing four key techniques, that is, spectro-temporal pattern matching, noise suppression using voice cancellation, beat prediction and music recognition. We implemented the proposed beat tracking method for Honda ASIMO. Experimental results showed 10-20 times faster adaptation to tempo changes and high robustness in beat tracking for stepping, scatting and singing noises. We also show some movies of our robot singer based on the beat-tracking.

1 はじめに

近年，ヒューマノイドやホームロボットなど人とソーシャルインタラクションを行うロボットの研究が盛んに行われている．実際に，ロボットが日常環境でソーシャルインタラクションを行うためには，自らの耳で音を聴き，それに応じて行動できる能力が必要である．音声コミュニケーション能力はいうまでもないが，音楽においても，人が演奏した音楽を聴いて，その音楽に応じて歌ったり踊ったりする機能は，人とのインタラクションをより自然で豊かにするために重要である．本稿では，音楽からビートを抽出し、これを基に人とのインタラクションを行う能力を有するロボットをビートトラッキングロボットと呼ぶものとする．ビートトラッキングロボット研究は，これまで主にソフトウェアを中心に行われてきた音楽情報処理研究に新たに物理的な身体を伴うエージェントやユーザインタフェースを提供するだけでなく，ビートトラッキングロボットによって音楽情報処理とロボティクスを融合した新たな研究領域の開拓が期待できよう．

1.1 ビートトラッキングロボットの課題と現状

リアルタイム・実環境で動作するビートトラッキングロボットの構築には，以下の課題が挙げられる．

(1) 音楽と挙動の同期
(2) 雑音ロバスト性
(3) リアルタイム処理
(4) テンポ変化に対応できるビートトラッキング
(5) 音楽認識

人・ロボットインタラクションのメディアとして，音楽への関心は高く，音楽に合わせて歌や踊りなどの挙動が生成できるロボットが複数報告されている．MIDI信号に合わせてダンスを行う WABIAN [1]，複数体で同期して，歌い踊ることができる Sony QRIO [2]，人と社交ダンスを踊る MSDanceR [Kosuge 03, Takeda 05, Takeda 06]，モーションキャプチャデータを利用して，踊りの模倣を行うロボット HRP-2 [Nakazawa 02] が挙げられる．これらのロボットは歌や踊りを行うことができるものの，事前にプログラムされた通りの動作を行っている．このため，人とのインタラクティブなセッションを行うような状況に不向きである．実際には，こうした歌や踊りは，音楽からリアルタイムで抽出したビート情報に基づき，音楽と挙動のテンポ，および音楽のビート時刻と挙動のタイミングの両方が同期するよう制御を行う必要がある．また，動作命令を発行してから，実際の動作までには遅延が生じるため，この遅延を予測しながら，上記の制御を行う必要がある．テンポが変化した場合のシステム全体のレスポンスは，テンポ変化検出までの遅延と実際に挙動に反映されるまで

[1] http://www.takanishi.mech.waseda.ac.jp/research/
[2] http://www.sony.net/SonyInfo/QRIO/

の遅延を合わせた遅延となるため，ロボットの制御においてもテンポ変化への高い追従性能が求められる．

実際に，ロボットのマイクから抽出した音楽情報を利用するロボットとして，Kotosaka &Schaal が開発した人とドラムを介してセッションを行うロボット [Kotosaka 00]，Michalowski らが開発した音楽ビートに合わせて，単純だが素早く愛らしい動作を行う小型ロボット Keepon [Michalowski 07] が挙げられる．しかし，これらのロボットは入力の音楽音響信号に雑音が含まれていない，もしくは，雑音が無視できる程度に入力音が大きいことを前提としている．一般的な環境では，雑音の混入はさけることができないため，周囲の環境雑音が大きい場合や，ロボット動作音など大きな自己発生音が生じる場合などは正確な動作が困難である．自己発生音の原因となるモータや発声用のスピーカは音楽音源よりもロボット搭載マイクに近い位置にあるため，相対的に雑音源のパワーが大きくなってしまう．また，音楽に合わせて動作や発声を行う場合，自己発生音自体が周期性を有するため，ロボットでビートトラッキングなどの処理を行う場合は自己発生音の抑制を行う必要がある．

明確にリアルタイム処理を謳っているビートトラッキング手法として後藤らのマルチエージェントベースのリアルタイムビートトラッキング [Goto 95] が挙げられる．この手法は，一定の処理遅れは生じるものの，リアルタイムでロバストにビートを抽出することができる．しかし，一定の処理遅れが生じるため，抽出したビート時刻は，理論的に過去の時刻のビートとなってしまう．このため，実ロボットに適用し，ビートに合わせた挙動生成を行うには，未来のビート時刻を予測する機能が必要である．また，より精度の高い予測を行うためには，処理遅れを極力抑える必要がある．さらに，ロボットでは，組込みを念頭に置き，スペースや電力消費にも配慮したリアルタイム処理を行うことが望ましい．

人の演奏のテンポは常に一定ではないこと，曲の途中でテンポが変わりうることから，ロボットを対象としたビートトラッキングには，テンポの変化への高い追従性能が求められる．一方で，MIDI 信号を用いる場合やビートが比較的一定である場合は，安定してビートが抽出できることが望ましい．安定的にビートを抽出するためには，ビート抽出の時間窓を長くするればよいが，テンポ変化への追従性は悪化してしまう．一般にテンポ変化への追従性と安定性はトレードオフ関係にあり，この両立は大きな課題である．従来は，テンポ変化を考慮していないビートトラッキングアルゴリズムが多かった [Goto 95, Scheirer 98] が，Dannenberg らは，Decay パラメータを用いて，このトレードオフを制御できる手法を考案している [Dannenberg 87]．また，近年では，パーティクルフィルタや確率的手法を導入して，テンポの揺れや変化への追従を考慮に入れた報告が見られる [Hainsworth 04, Klapuri 06, Cemgil 03]．パーティクルフィルタは，パーティクル数・処理速度と精度がトレードオフ関係にあり，確率モデルは，事前に確率分布を計算する必要があるといった欠点がある．

演奏された音楽に応じた挙動を生成するためには，その音楽の曲名や，歌詞，歌い始めのタイミングといった情報が必要である．このためには，ビートトラッキングだけではなく，音楽の認識および，認識した音楽に対する知識の検索機能が必要である．音楽の認識に関しては，ISMIR (International Conference on Music Information Retreival) に代表されるように，音楽情報検索研究の分野で盛んに行われており，これらの知見を利用することが可能であろう．

1.2 課題解決のアプローチ

本稿では，上記の課題に対し，下記の 4 つの手法の導入を試みる．

- 音楽と挙動のテンポとビート時刻のずれを最小化するフィードバック制御の導入（課題 (1) への対応）．
- 自己発生音を抑制するため，セミブラインド独立成分分析 (セミブラインド ICA)[Takeda 07] の導入（課題 (2) への対応）．
- 時間周波数領域でパターンマッチングを行うことにより，一般的な自己相関関数を用いる場合よりも短い窓長で，ロバストにビートを抽出することができる高速なビートトラッキング手法である STPM (Spectro-Temporal Pattern Maching) ビートトラッキング法の提案（課題 (3),(4) への対応）．
- 音楽区間検出と曲名検索機能の導入 (課題 (5) への対応)．

さらに，Honda ASIMO に対して，提案手法を実際に適用し，ロボット自身に備えられたマイクを用いて抽出したビートに合わせて，足踏みを行い，同時に口ずさんだり，歌声を出力したりできるビートトラッキングロボットを報告する．

2 STPM ベースのビートトラッキング

上記の方針を元に提案するビートトラッキングアルゴリズムの概要を Fig. 1 に表す．このビートトラッキングアルゴリズムも「周波数解析」「テンポ予測」「ビート時刻予測」の 3 つのモジュールからなっており，入力信号に対してビート時刻とテンポを出力する．

2.1 周波数解析

44.1kHz で同期してサンプリングした 2ch の入力信号に対して周波数解析を行う．1 チャンネルはロボットに内蔵されたマイクを使用し録音した音楽音響信号である．そのため，音楽音響信号の他に環境ノイズや自己発生音などの雑音が混入している．もう 1 チャンネルは口ずさみや

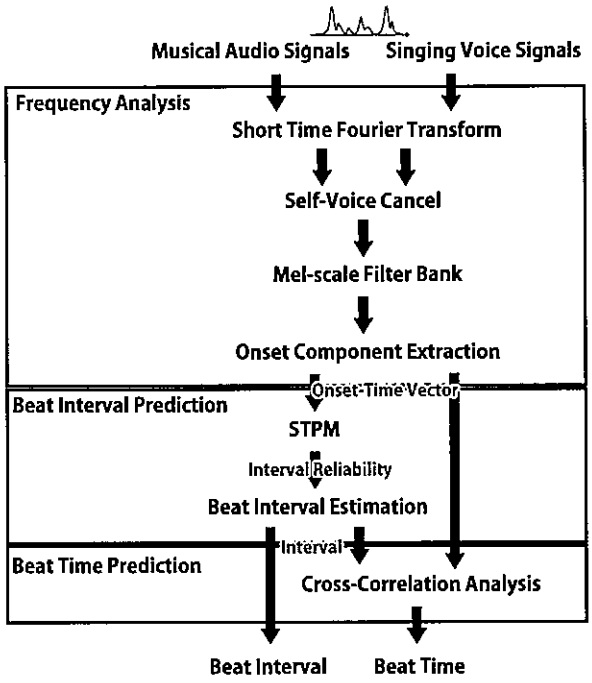


Figure 1: SPTM ビートトラッキングアルゴリズム概要

歌声などのロボットの自己発声音の信号である．システムが生成する信号を直接入力しているため，雑音が混入していない．この信号は後述する雑音抑制で用いる．

周波数解析では2つの信号それぞれに対して短時間フーリエ変換(STFT)を行う．ここで，窓関数には窓長4096ポイントのハニング窓を用い，シフト長は512ポイントとした．

次に口ずさみや歌声など音楽に応じて行われる周期性のある自己発声音のキャンセルを行う．ビートトラッキングは音楽の周期性を用いてビートを検出するため，発声の影響が大きい．このため，自己発声音を抑制することによってビートトラッキングの精度向上が期待できる．具体的な雑音除去にはセミブラインドICA[Takeda 07]に基づくボイスキャンセル法を使用する．この手法は一般的な適応フィルタベースの手法より性能が高いことが報告されておりマルチチャンネル入力にも簡単に拡張できるという特徴を備えている．本稿では前述の2ch入力信号に対し，この手法を適用した．次に，スペクトルに対し音声認識や音楽認識で用いられるメルフィルタバンクを用いて2,049次元のスペクトルを64次元に圧縮した．

メルスケールのスペクトログラム上でパワーが急激に上昇している時刻はオンセットである可能性が高いと考える．まず定常雑音を除去するため，パワースペクトログラムに対してソーベルフィルターを適応してエッジの強調を行う．$p_{mel}(t,f)$ を t 番目の時間フレームの f 番目のメル周波数ビンのスペクトルパワー値とすると，フィルター後の値は $d(t,f)$ と表すことが出来る．

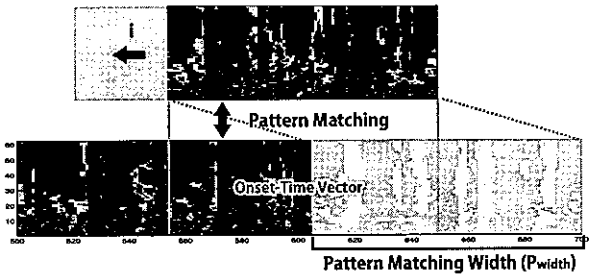


Figure 2: 周波数ー時間　パターンマッチング．

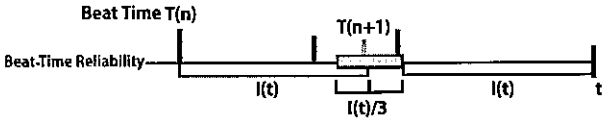


Figure 3: ビート時刻推定

$$
\begin{aligned}
d(t,f) = \quad & -p_{mel}(t-1,f+1)+p_{mel}(t+1,f+1) \\
& -p_{mel}(t-1,f-1)+p_{mel}(t+1,f-1) \\
& -2p_{mel}(t-1,f)+2p_{mel}(t+1,f) \qquad (1)
\end{aligned}
$$

このとき，オンセット信頼度 $d_{inc}(t,f)$ を以下のように定義する．

$$
d_{inc}(t,f) = \begin{cases} d(t,f) & \text{if d(t,f)} > 0, \\ 0 & \text{otherwise} \end{cases} \qquad (2)
$$

ここで $f = 1,2,...,62$ である．よってそれぞれの時間フレームで62次元のオンセット時刻ベクトルが得られる．

2.2 ビート間隔推定

隣り合う二つのビートの間隔を「ビート間隔」と定義し，これを推定する．まず，オンセット時刻ベクトルを用いて時間-周波数領域でパターンマッチングを行う．パターンマッチング関数として以下で定義される正規化相互相関関数(NCC)を用いる．ビート間隔信頼度 $R(t,i)$ を

$$
\begin{aligned}
R(t,i) &= \frac{A(t,i)}{\sqrt{B(t)C(t,i)}} \qquad (3) \\
A(t,i) &= \sum_{j=1}^{62}\sum_{k=0}^{P_{width}-1} d_{inc}(t-k,j)d_{inc}(t-i-k,j) \\
B(t) &= \sum_{j=1}^{62}\sum_{k=0}^{P_{width}-1} d_{inc}(t-k,j)^2 \\
C(t,i) &= \sum_{j=1}^{62}\sum_{k=0}^{P_{width}-1} d_{inc}(t-i-k,j)^2
\end{aligned}
$$

と定義する．ここで P_{width} はパターンマッチングの窓長で i はシフトパラメータである．(Fig. 2).

提案するアルゴリズムは時間-周波数パターンマッチングを用いているため時間と周波数の情報を同時に利用し

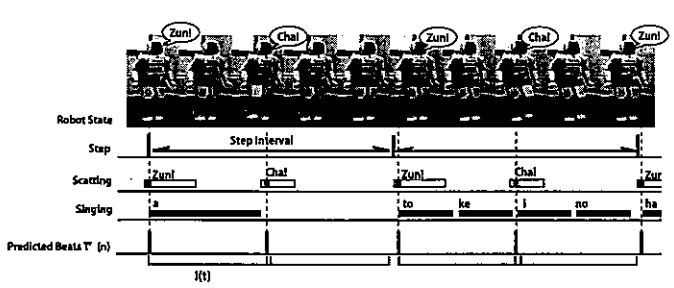


Figure 4: ビートトラッキングロボットの動作

ている．このため窓長が1秒程度と短い場合でもロバスト性を保ったままテンポ変化に対して素早い追従を行うことが出来る．これは，例えば，後藤らのアルゴリズム[Goto 95]で，自己相関関数の窓長が約6〜10秒であることからも短いといえよう．

次に $R(t,i)$ からローカルピークを抜き出す．

$$R_{peak}(t,i) = \begin{cases} R(t,i) & \text{if R(t,i-1)} < \text{R(t,i)} < \text{R(t,i+1)}, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

時刻 t でのビート間隔の候補として，$R_{peak}(t,i)$ を最大化する i と二番目に大きくする i をそれぞれ $I_1(t), I_2(t)$ とおく．この2つのピークの信頼度が同程度であるとき，一方は偽のビート間隔を検出していると考えられる．偽のビート間隔の検出を避けるために，まずビート間隔をロボットの動作範囲である61M.M. から120M.M.[3] に制限する．裏拍などの影響で2倍3倍のビート間隔が間違いとして検出されやすいことを考慮して，本来のビート間隔を以下のようなヒューリスティクスを用いて定義する．

$$I_{new} = \begin{cases} 2|I_1 - I_2| & (|I_{n2} - I_1| < \delta \text{ or } |\text{I}_{n2} - \text{I}_2| < \delta) \\ 3|I_1 - I_2| & (|I_{n3} - I_1| < \delta \text{ or } |\text{I}_{n3} - \text{I}_2| < \delta) \end{cases} \tag{5}$$

$$I_{n2} = 2|I_1 - I_2|$$

$$I_{n3} = 3|I_1 - I_2|$$

ここで δ は許容誤差である．

I_{new} が存在するとき，ビート間隔 $I(t)$ は I_{new} にセットされ，存在しなければ I_1 と I_2 の信頼度の高い方が $I(t)$ となる．

2.3 ビート時刻予測

ビート時刻信頼度は近接ビート信頼度と連続ビート信頼度の二種類から算出される．

2.3.1 近接ビート信頼度

近接ビート信頼度は，ある時刻とそのビート間隔前の時刻がともにビート時刻である信頼度を表す．時刻 t における時刻 $t-i$ の近接ビート信頼度 $S_c(t,i)$ は，時刻 $t-i$ と

[3] Mälzel's Metronome: 一分あたりの四分音符の数．たとえばテンポ60 M.M. であれば，四分音符の長さは1,000 [ms] である．

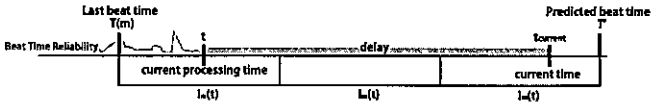


Figure 5: 未来ビート時刻予測概要

そのビート間隔前の時刻 $t-i-I(t)$ のオンセット時刻ベクトルを用いては以下のように表せる．

$$S_c(t,i) = \begin{cases} F_s(t-i) + F_s(t-i-I(t)) & (i \le I(t)) \\ 0 & (i > I(t)) \end{cases} \tag{6}$$

$$F_s(t) = \sum_{f=1}^{62} d_{inc}(t,f)$$

2.3.2 連続ビート信頼度

連続ビート信頼度とはビート間隔でビートが連続して存在している信頼度である．近接ビート信頼度を用いて以下のように定義した．

$$S_r(t,i) = \sum_{m}^{N_{S_r}} S_c(T_p(t,m),i) \tag{7}$$

$$T_p = \begin{cases} t - I(t) & (m=0) \\ T_p(t,m-1) - I(T_p(t,m)) & (m \ge 1) \end{cases}$$

ここで $S_r(t,i)$ は時刻 t における時刻 $t-i$ に存在するビートに対する連続ビート信頼度を意味する．$T_p(t,m)$ は時刻 t を基準として m 個前のビート時刻で，N_{S_r} は連続ビート信頼度を評価する際のビート数である．この信頼度は複数のビート列が見つかった場合に，一番強いビート列を求める際に用いる．

2.3.3 ビート信頼度

近接ビート信頼度と連続ビート信頼度を用いてビート信頼度を以下のように定義する．

$$S(t) = \sum_i S_c(t-i,i) S_r(t-i,i) \tag{8}$$

2.3.4 ビート時刻検出

まず，n 番目のビート時刻を $T(n)$ とする．$T(n) \ge t - \frac{3}{4}I(t)$ の時，区間 $[T(n)+\frac{1}{2}I(t), T(n)+\frac{3}{2}I(t)]$ 内で上位3つのピークを抽出する．次にFig. 3に示されるように，$T(n)+I(t)$ に一番近いピークを次のビート時刻 $T(n+1)$ とする．ピークが区間 $[T(n)+\frac{2}{3}I(t), T(n)+\frac{4}{3}I(t)]$ に存在しない場合，$T(n)+I(t)$ を $T(n+1)$ とする．

2.3.5 未来のビート時刻予測

Fig. 5のように計算での最新時刻 t は計算時間や歌声生成のため現在時刻 $t_{current}$ に対して遅れが存在する．よって最新のビート時刻 $T(m)$ は現在時刻 $t_{current}$ より前の時刻のビート，つまり $t_{current} > T(m)$ である．口ずさみや歌機

能を実現するためには未来のビート時刻 T' を予測する必要がある．今回は単純に以下の外挿を用いることで，未来のビート時刻を予測する．

$$T' = \begin{cases} T_{\mathrm{tmp}} & \text{if } \mathrm{T_{tmp}} \geq \frac{3}{2}\mathrm{I_m(t)} + \mathrm{t} \\ T_{\mathrm{tmp}} + I_m(t) & \text{otherwise.} \end{cases} \quad (9)$$

$$T_{\mathrm{tmp}} = T(m) + I_m(t) + (t - T(m)) - \{(t - T(m)) \mod I_m(t)\}$$

ここで $I_m(t)$ は $I(t)$ 群の中央値で，$T(m)$ は検出したビート時刻の中で一番新しい時刻である．

3 ビートトラッキングロボットの実装

Fig. 6 に我々のビートトラッキングロボットのアーキテクチャを示す．このシステムは主にロボットと三つのサブシステム (リアルタイムビートトラッカー，音楽認識，ロボット制御) に分けることができる．

3.1 ロボット

頭部にマイクロフォン（1本）を搭載したホンダ ASIMO を用いた．足踏み間隔は 1,000 から 2,000 [ms] の範囲となっている．これは音楽のテンポに換算すると 61〜 120 M.M. である．胸部には発声用にスピーカーを内蔵している．

3.2 リアルタイムビートトラッカー

リアルタイムビートトラッカーはロボットのマイクロフォンから入力された音楽音響信号を用いてビート間隔とビート時刻を計算し出力する．このとき口ずさみや歌声といった出力信号を同時に入力することで音楽音響信号に混入した自己発声音をキャンセルする．詳細なアルゴリズムは 2 章を参照されたい．

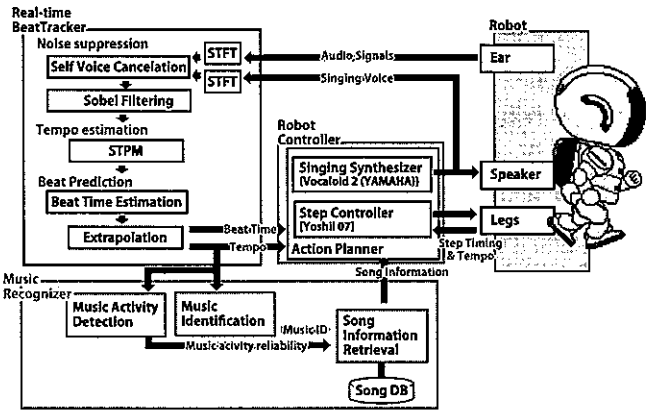


Figure 6: ビートトラッキングロボットの構成図

3.3 音楽認識 (Music Recognizer)

音楽認識は音楽区間検出と曲名検索からなる．音楽区間検出では，音楽は安定したビート間隔をもっていると仮定する．ビート安定度として過去 3 秒間にある現在の時刻のビート間隔と近いビート間隔を持つ時間の比を用いる．ビート間隔の差が 55ms 以下であれば近いとする．この安定度が 0.8 以上の時音楽が存在すると見なす．この値は経験的に決めた．

曲名検索では予測したビートと近いビートを持つ曲を曲データベースから検索し，曲 ID を返す．今回ビートの近さの指標にはシンプルにビートのテンポの情報のみを用いた．経験的に予測したテンポと曲のテンポの差が 11ms 以内であった場合，同じ曲だと見なした．その曲 ID を用いて曲情報データベースから歌詞情報と譜面情報をえる．もし，曲が存在しなければ「unknown」を曲 ID として返す．この場合は「ずん」「ちゃ」という声を用いて口ずさみを行う．これらの情報はロボット制御へと送られる．

3.4 ロボット制御 (Robot Controller)

ロボット制御はリアルタイムビートトラッカーによって検出されたビート時刻とビート間隔、音楽認識によって検索された曲情報にあわせて，同期して足踏みしたり，口ずさんだり歌を歌うようにする．Fig. 4 はビートトラッキングロボットの動きの例を表している．

3.4.1 足踏み

今回のロボット制御で使うことが出来るのは足踏み間隔を変化させる命令のみだったので，これを用いてロボットの足踏み間隔と時刻を抽出したビート時刻と間隔に合わせる．典型的なフィードバック制御理論では出力に誤差があっても許容されるが，目標値の誤差は考慮されていないため予測誤差や量子化誤差を含んでいる場合など，一般に正確な目標値の設定が困難な場合，フィードバック制御は難しい．そこで，下記の方法で足踏みとビートの間隔と時刻の差を同時に低減した [Yoshii 07].

$$\begin{aligned} I_{in}(n+1) &= I_{in}(n) + \beta_I \left(I(n) - I_{out}(n)\right)) \\ &\quad + \beta_T \left(T(n) - T_{out}(n)\right)), \end{aligned} \quad (10)$$

ここで $I_{in}(n)$ は n 番目のロボットへの足踏み間隔指示，$I(n)$ がシステムの推定したビート間隔，$I_{out}(n)$ が実際のロボットの足踏み間隔，$T(n)$ がシステムの推定したビート時刻，$T_{out}(n)$ が実際のロボットの足踏み時刻，β_I と β_T は，足踏み間隔と時刻のどちらを優先して同期するかを調節する重みである．二つの重み付けの要素である β_I と β_T の値を以下の閾値で変化させる．

$$|I(n) - I_{out}(n)| < \varepsilon I(n), \quad (11)$$

ここで，ε は許容誤差で，ここでは小さな値 0.02 とする．足踏み間隔の誤差が大きいとき，つまり式 11 を満たさないとき，β_I と β_T は経験的に 0.30 と 0.00 にセットされる．つまり間隔の誤差が大きいときは間隔の誤差を減らすことを優先する。もし，式 11 が満たされたら，β_I と β_T は 0.10 と 0.02 にセットされる．

3.4.2 口ずさみ

口ずさみは，ビート時刻に合わせて「ずん」「ちゃ」と口ずさむ機能である．ビート時刻に合わせて自然に聞こえるよう口ずさむためには「ずん」や「ちゃ」の各音とビート時刻の同期のタイミングが重要である．そこで「ずん」，「ちゃ」から抽出したオンセット時刻ベクトルの各値の合計値のピークを「ずん」「ちゃ」のビート時刻とする．この各音でのビート時刻と，音楽のビート時刻を合わせて発音する．

3.4.3 歌唱

歌唱機能は，聴いた曲の速度とビートに同期して歌を歌うという機能である．今回はビート時刻ごとに音符の発音時刻と発音長 (duration)，発音記号をあらかじめ入力した楽譜（音階）と歌詞から計算し，MIDI信号を出力する．このMIDI信号をVOCALOID 2に送出することによって歌唱機能を実装する．VOCALOID 2はMIDIを用いて自然な歌声を合成できる商用ソフトである．VOCALOID 2はMIDIデータを受け取ってから発音するまでに200[*msec*]の遅れがある．それを考慮し，発音時刻の200[*msec*]前にMIDIデータを送信するよう実装を行った．

4 実験

ビートトラッキングロボットを以下の3点で評価した．

実験1 テンポ変化への追従速度
実験2 ビート予測のノイズロバスト性能
実験3 ノイズ下での音楽認識性能

これらを評価するために以下の3種類の音楽信号を用いて実験を行った．

T1 テンポ変化を含む音楽音響信号
RWC 音楽データベース (RWC-MDB-P-2001)[Goto 02] から3曲 (No. 11, No. 18, No. 62) を選んだ．これらは市販のCDのように様々な楽器音と歌声を含んでおり，テンポはそれぞれ90, 112, 81 M.Mである．これらをNo.18 - No.11 - No.18 - No.62の順に60秒ずつ区切りでつなげることで4分の音楽信号を作成した．

T2 テンポの固定された音楽音響信号
No. 62のMIDIデータを用いて生成したテンポ一定の曲を用いる．ただし，MIDIのデータはビート時刻の検証のみに用いる．

T3 ノイズ下での音楽音響信号
RWC 音楽データベースから5曲 (No. 4, No. 11, No. 17, No. 18, No. 29) を含む10分のデータを用意した．20秒ごとにノイズのみの区間と曲にノイズが混じった区間が入れ替わる．ノイズにはJDEIDA-NOISEデータベースの「展示会場でのブース内の音」を用いた．SNRは平均約-4dBである．

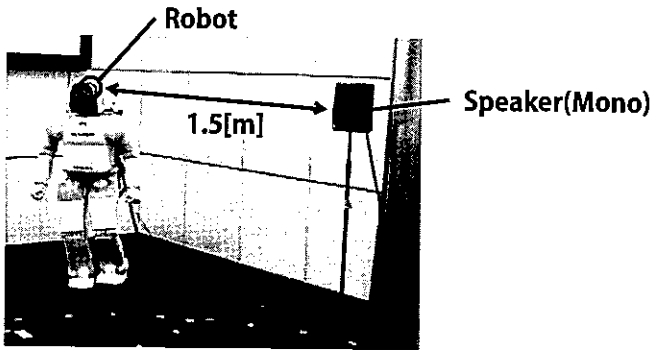


Figure 7: 実験環境

実験はすべて残響時間0.2秒 (RT_{20})，4 m×7 mの部屋で行い，音楽用音源にはスピーカー (GENELEC 1029A) を用いた．ロボットとスピーカの距離は1.5 mである（Fig. 7）．

まず**T1**を用いて5種類の状態でビートトラッキングの遅れを計測した．ビートトラッキングの遅れは実際にテンポが変化してからシステムがテンポ変化に追従するまでの時間とする．適応した時間の差とする．5種類のうち2種類はASIMOの電源をoffにして口ずさみを行った場合と行わなかった場合，他の3種類はASIMOの電源をonにして足踏みをしながら口ずさみを行った場合，行わなかった場合，歌った場合である．

実験2では**T2**を用いて5種類の状態でビート推定の成功率を計測した．ビート推定の成功率rは以下のように定義する．

$$r = \frac{N_{\text{success}}}{N_{\text{total}}} \times 100. \tag{12}$$

ここでN_{success}は推定が成功したビート数で，N_{total}は正解ビート総数である．推定されたビート時刻と正解ビート時刻の差が$\pm 0.35I(t)$以内に収まっている場合は推定が成功したものとした．5種類のうち3種類はASIMOの電源を切った状態で，1種類は口ずさみなしでボイスキャンセルあり，他の2種類は共に口ずさみありだが，ボイスキャンセルの有無に違いがある．他の2種類はASIMOの電源を入れ足踏みをしながら口ずさみを行っている状態でボイスキャンセルを行った場合と行わなかった場合である．

実験3では**T3**を用いて5曲それぞれの認識率を測定した．我々は指標として適合率 (P)，再現率 (R)，F値 (F) を以下のように定義した．

$$P = \frac{C}{N}, \quad R = \frac{C}{A}, \quad F = \frac{2 \cdot P \cdot R}{P + R} \tag{13}$$

ここでCは音楽のある区間が正しく認識出来た時間を表し，Nは音楽がある区間であると認識した時間を表し，A実際の音楽の時間を表す．音楽認識の指針として、音楽認識率 (M) を以下のように定義した。

$$M = \frac{C'}{N} \tag{14}$$

ここでC'音楽が正しく認識された総時間を表す．

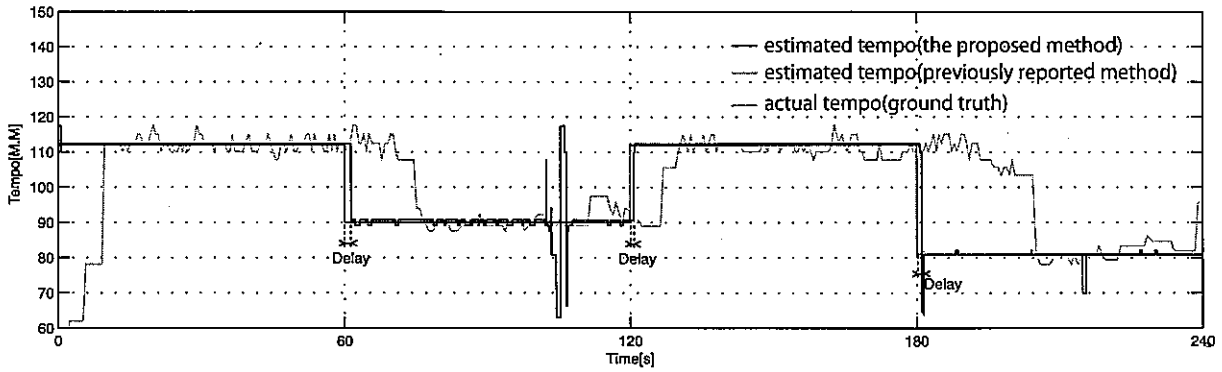


Figure 8: テンポ変化のある音楽を用いた結果

Table 1: テンポ変化に対する追従遅延

	ASIMO power off		ASIMO power on(with step)		
	w/o scatting	w/ scatting	w/o scatting	w/ scatting	w/ singing
reported	11.24	29.91	14.66	20.43	N/A
proposed	1.31	1.31	1.29	1.29	1.29

Table 2: ビート予測正解率

	ASIMO power off			ASIMO power on (with step)	
	w/o scatting	w/ scatting		w/ scatting	
		w/ echo cancel	w/o echo cancel	w/ echo cancel	w/o echo cancel
Correct	95%	97%	68%	95%	64%
Half shifted	5%	1%	40%	4%	40%

4.1 結果

Fig. 8 は実験 1 の結果を表している．提案手法は従来手法に比べテンポ変化への適応が速いことがわかる．100 秒近くでビートトラッキングが乱れているのは T1 がビート時刻にオンセットがない部分が一時的に存在しているからである．このため提案手法では一時的に，従来手法では長期に乱れる．遅れの平均値を Table 1 に示す．我々の提案した手法が従来手法より口ずさみのない場合で 10 倍程度，ある場合で 20 倍程度追従が高速であることがわかる．

Table 2 は実験 2 の結果を表している．“正解” はビートトラッキングシステムが正しいビートを予測した事を示し，“半分ずれ” は裏拍を予測した事を示す．これは自己発声音がその周期性のためにビートトラッキングに影響を与えていること，およびボイスキャンセルが影響をこうしたノイズに効果的に働くことを示している．

また実際に口ずさみや歌いながら音楽のビートに合わせてロボットが足踏みが出来ることも確かめられた．

Table 4.1 は実験 3 の結果を表す．適合率の平均は再現率の平均より 10 ポイント程度高い．これは音楽区間検出に 3 秒間の窓を用いているため初めの 2.4 秒間 (3×0.8) が検出不可能であるからである．ノイズが 90〜97 bpm 付近の周期音を含んでいるため，曲#11 と #17 はノイズの影響を受けて再現率が落ちている．M はクリーンデータで 95.8%，ノイズを含んだデータでは 88.5% だった．これらから音楽認識はテンポ情報しか使っていないにもかかわらず，少ない数の曲に関してはよい性能をもっていると言える．音楽認識のスケーラビリティを良くするためには，たとえばリズム特徴 [Kirovski 02] のようなもっと高度な情報を使う必要がある．

Table 3: 音楽認識

ID	bpm	with noise			clean		
		P (%)	R (%)	F	P (%)	R (%)	F
#4	86	94.7	84.9	0.90	94.8	81.2	0.87
#11	90	74.3	67.3	0.71	96.1	72.1	0.82
#17	97	88.0	83.1	0.85	95.3	81.6	0.88
#29	103	93.4	81.5	0.87	95.9	82.2	0.88
#18	112	89.6	82.8	0.86	95.9	83.2	0.89

5 おわりに

ビートトラッキングロボットを実現するために，リアルタイム性・安定性・テンポ変化への追従性を備えた時間周波数パターンマッチングベースのビートトラッキング手法を提案した．また，口ずさみや歌などの周期性を伴った自己発声音の影響を削減するためにセミブラインド ICA によるボイスキャンセルを行った．さらに提案した手法を Honda ASIMO に実装し，ロボット搭載マイクを用いて収録した信号からリアルタイムでビートを抽出し，そのビート情報に基づき，足踏み，口ずさみ，歌唱を行うビートトラッキングロボットを構築した．この際，フィードバック制御手法を導入し音楽と挙動のずれを最小化を行った．構築したシステムを用いて提案手法を評価し，ノイズロバスト性とテンポ変化に対する高追従性を確認した．ロボットの挙動の高度化，ロボット外の雑音源の抑制，画像など他のモダリティを用いたロバスト性の向上が今後の課題である．

参考文献

[Cemgil 03] Cemgil, A. and Kappen, B.: Monte Carlo Methods for Tempo Tracking and Rhythm Quantization., *Journal of Artificial Intelligence Research*, Vol. 18, pp. 45–81 (2003)

[Dannenberg 87] Dannenberg, R. and Mont-Reynaud, B.: Following an Improvisation in Real Time., in *Proceedings of the International Computer Music Conference*, pp. 241–258, International Computer Music Association (1987)

[Goto 95] Goto, M. and Muraoka, Y.: A Real-Time Beat Tracking System for Audio Signals., in *Proceedings of the International Computer Music Conference*, pp. 171–174, San Francisco CA (1995), International Computer Music Association

[Goto 02] Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases, in *Int. Conf. Music Information Retrieval*, pp. 287–288 (2002)

[Hainsworth 04] Hainsworth, S. W. and Macleod, M. D.: Particle Filtering Applied to Musical Tempo Tracking, *EURASIP Journal on Applied Signal Processing*, Vol. 15, pp. 2385–2395 (2004)

[Kirovski 02] Kirovski, D. and Attias, H.: Beat-ID: Identifying Music via Beat Analysis, in *IEEE Workshop on Multimedia Signal Processing*, pp. 190–193 (2002)

[Klapuri 06] Klapuri, A. P., Eronen, A. J., and Astola, J. T.: Analysis of the Meter of Acoustic Musical Signals, *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 14(1), (2006)

[Kosuge 03] Kosuge, K., Hayashi, T., Hirata, Y., and Tobiyama, R.: Dance Partner Robot -MS DanceR-, in *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS-2003)*, pp. 1743–1750 (2003)

[Kotosaka 00] Kotosaka, S. and Schaal, S.: Synchronized Robot Drumming by Neural Oscillators, in *Proc. of Int'l Sympo. Adaptive Motion of Animals and Machines* (2000)

[Michalowski 07] Michalowski, M., Sabanovic, S., and Kozima, H.: A dancing robot for rhythmic social interaction, in *Proc. of ACM/IEEE Int'l Conf. on Human-Robot Interaction (HRI 2007)*, pp. 89–96, IEEE (2007)

[Nakazawa 02] Nakazawa, A., Nakaoka, S., Ikeuchi, K., and Yokoi, K.: Imitating Human Dance Motions through Motion Structure Analysis, in *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robot s and Systems (IROS-2002)*, pp. 2539–2544 (2002)

[Scheirer 98] Scheirer, E.: Tempo and Beat Analysis of Acoustic Musical Signals., *Journal of the Acoustical Society of America*, Vol. 103(1), pp. 588–601 (1998)

[Takeda 05] Takeda, T., Hirata, Y., and Kosuge, K.: HMM-Based Dance Step Estimation for Dance Partner Robot - MS DanceR-, in *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS-2005)*, pp. 1602–1607 (2005)

[Takeda 06] Takeda, T., Hirata, Y., Wang, Z., and Kosuge, K.: HMM-based Error Detection of Dance Step Selection for Dance Partner Robot –MS DanceR–, in *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS-2006)*, pp. 5631–5636 (2006)

[Takeda 07] Takeda, R., Nakadai, K., Komatani, K., Ogata, T., and Okuno, H. G.: Exploiting Known Sound Sources to Improve ICA-based Robot Audition in Speech Separation and Recognition, in *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS-2007)*, pp. 1757–1762 (2007)

[Yoshii 07] Yoshii, K., Nakadai, K., Torii, T., Hasegawa, Y., Tsujino, H., Komatani, K., Ogata, T., and Okuno, H. G.: A Biped Robot that Keeps Steps in Time with Musical Beats while Listening to Music with Its Own Ears, in *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS-2007)*, pp. 1743–1750 (2007)

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A802-4 (11/18)

Robust Speech Recognition in Reverberant Environment by Optimizing Multi-band Spectral Subtraction

Randy Gomez and Tatsuya Kawahara
Kyoto University, Academic Center for Computing and Media Srudies (ACCMS),
Sakyo-ku, Kyoto 606-8501, JAPAN

Abstract

Reverberant environment poses a problem in speech recognition application where performance degrades drastically depending on the extent of reverberation. Thus, it is important to employ front-end speech processing, such as dereverberation to minimize its effect. Most dereverberation techniques used to address this problem enhance the reverberant waveform prior to speech recognition. Although the speech quality is improved, this approach treats the front-end speech enhancement and the recognizer independently. In this paper, we present an approach that treats both dereverberation and speech recognition interdependently. In our proposed approach, the dereverberation parameters are optimized to improve the likelihood of the acoustic model. The system is capable of adaptively fine-tuning these parameters jointly with acoustic model training. Additional optimization is also implemented during decoding of the test utterances. Experimental results show that the proposed method significantly improves the recognition performance over the conventional approach with a relative improvement of 5%.

1 Introduction

In hands-free speech recognition applications, the observed speech signal at the microphone is smeared by a phenomenon known as reverberation. This is due to the reflection of the speech signal inside a closed space (i.e. room). The smearing varies significantly with the property and dimension of the room. The recognition performance of a reverberant test utterance using a reverberant model is significantly degraded compared to the performance of non-reverberant test utterance with a non-reverberant model. Thus, it is imperative to counter the negative effect of reverberation both the test data and the acoustic model.

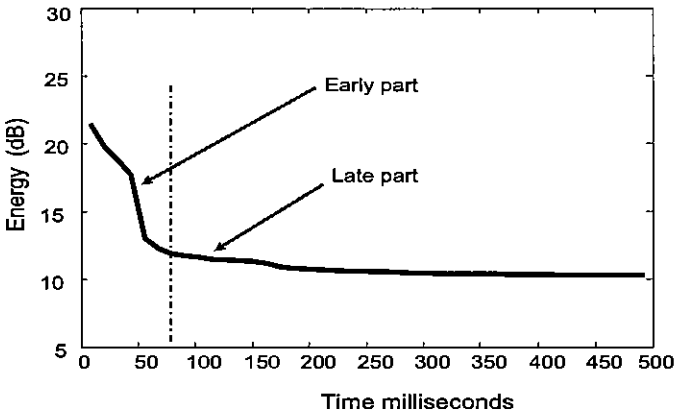


Figure 1: Measured impulse response energy.

We have proposed a single channel framework dereverberation technique based on multi-band Spectral Subtraction (SS) [1][2]. Similar approach based on single-band SS has been proposed in the work of [3]. In the multi-band SS dereverberation technique, the late reflection of the observed reverberant signal is suppressed through multi-band SS, whereas the early reverberant part (early reflection), more likely to vary with microphone-speaker distance, is handled through Cepstrum Mean Normalization (CMN) [4] [5]. The extent of suppressing the effects of the late reverberant signal is a function of the multi-band coefficients which are optimized using Minimum Mean Square Error (MMSE) criterion. Although this scheme works well, this criterion is inclined in optimizing the effect of dereverberation in the waveform level. Typically, this is a speech enhancement approach which improves the quality of the signal prior to acoustic modeling and recognition. This set-up treats the speech enhancement and recognition independently.

In this paper, we propose to treat these two interdependently by optimizing the dereverberation parameters based on the speech recognizer. Instead of just using the MMSE, we modified the criterion to directly optimize the likelihood of the recognizer. In this paper,

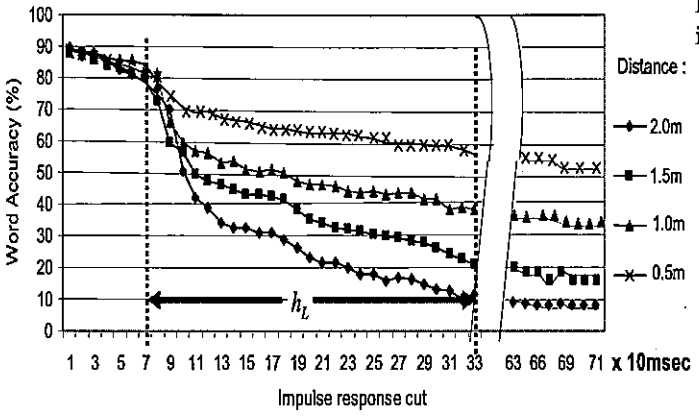


Figure 2: Late reflection boundary identification using recognition experiments and the measured room impulse response.

the optimization process of the dereverberation parameters is embedded in the acoustic model training. As a result, the dereverberation parameters are updated together with the acoustic model. This kind of approach, where front-end speech processing is optimized for recognition is shown to be effective with promising results in microphone array applications [6][7] and in Vocal Tract Length Normalization (VTLN) [8][9][10].

The organization of the paper is as follows; in section 2, we discuss the background of reverberation including its mathematical model as a function of early and late reflection. We also discuss the concept of multi-band SS based on the MMSE criterion as a dereverberation scheme. In section 3, we present the optimization in the acoustic model training phase. This involves optimization of the multi-band SS parameters based on the likelihood. In section 4, the optimization during decoding is presented. Experimental results are given in section 5, and we will conclude this paper in section 6.

2 Dereverberation Scheme

In this section, we discuss the significance of the room impulse response and its effect in the context of early and late reflection. In addition, we explain its characteristics relative to the Hidden Markov Model (HMM) structure. Consequently, we present the mathematical concept of multi-band Spectral Subtraction as a derverberation technique used in suppressing the effects of the late reflection.

2.1 Reverberation and Impulse Response

A reverberant speech signal contains the effects due to the early and late reflection. Room impulse response gives a good insight of reverberation and is often used to experimentally create a reverberant speech. When referring to the early reflection, we include by definition the direct speech signal and the overlapping of speech at earlier time. The late reflection however, is the collective overlapping of reflected speech at much later time. The following are the characteristics of the early and late reflection based on the energy plot of the measured impulse response $h(n)$shown in Figure 1:

(1) Early reflection has higher energy compared to the late reflection. Thus the speech signal in this region is dominant.

(2) Early reflection has a more dynamic value as compared to the late reflection which tend to be static over time. This characteristic implies that the effect of the late reflection can be approximately treated as constant. Since late reflection is a result of the overlapping of the speech signal in a much later time, a static energy means that as the distance between the speaker and the microphone increases, the characteristic of the late reflection remains relatively the same. Hence, a single impulse response measurement is enough to represent the different microphone-speaker locations. This treatment cannot be applied to the early reflection as its dynamic nature suggests that is sensitive to microphone-to-speaker locations.

(3) When considering a 3-state HMM architecture which has a 25 msec window and 10 msec window period, the early reflection occurs within the HMM architecture is designed to handle. Whereas, late reflection falls outside of the analysis framework.

Based on the arguments above, it is reasonable to argue that it would be beneficial to remove only the effect of late reflection through signal processing (i.e. using Spectral Subtraction) and retain the effect of the early reflection. The latter is more dependent with speaker-microphone distance, thus removing it together with the late reflection would require different impulse response measurement depending on the different microphone-speaker locations. In addition, the early reflection can be handled by the model-based system (HMM) through Cepstral Mean Normalization [4] [5].

2.2 Spectral Subtraction-based Dereverberation

In this section we outline the conventional dereverberation technique based on multi-band SS [1][2]. The speech signal has a strong correlation within each local time frame due to articulatory constraints. However, this correlation is lost according to articulatory movements [3]. As a result, it is established that early and late reflection are uncorrelated. Thus the reverberant speech signal $x(n)$ can be modeled as

$$x(n) = x_E(n) + x_L(n), \tag{1}$$

where $x_E(n)$, $x_L(n)$ are the uncorrelated early and late reflection components of the reverberant signal $x(n)$. If we denote $s(n)$ as clean speech, and the measured room impulse as $h(n) = [h_E(n), h_L(n)]$ where early components $h_E(n)$ and late components $h_L(n)$ of the whole sample $h(n)$ are identified in advance, Eq (1) can be written as,

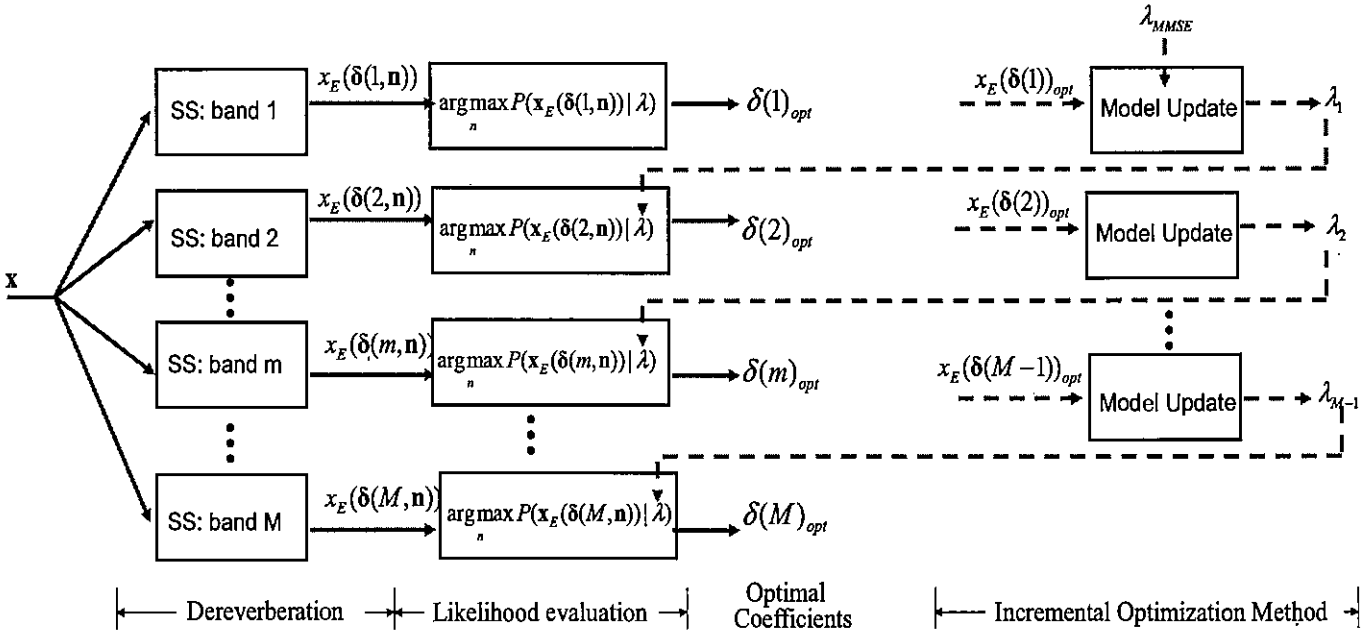


Figure 3: Block diagram of the proposed optimization technique in the acoustic training phase.

$$x(n) = h_E * s(n) + h_L * s(n). \quad (2)$$

The boundary of the early and late reflection is very important in our model. Figure 2 is used in identifying the said boundary, where the horizontal axis represents the length of the impulse response and the vertical axis shows the recognition performance. It is obvious in this figure that the steep decrease in the performance starts at 70 ms which suggests the beginning of the effect of the late reflection. The steep decrease is attributed to the fact that the recognizer cannot deal with reverberation that fall outside of the 3-state HMM structure (i.e. caused by $x_L(n)$). Moreover, the insignificant decrease in the recognition performance within 70msec suggest that the recognizer can handle the effect due to $x_E(n)$.

In the SS-based dereverberation, we are only interested in recovering $x_E(n)$ from $x(n)$. Thus, we use spectral subtraction to remove the effect of $x_L(n)$. Theoretically, it is possible to remove entirely the effect of the whole impulse response $h(n)$, but robustness to the microphone-speaker location cannot be achieved since the early components $h_E(n)$ have high energy and is dependent on the distance between the microphone and speaker as explained in [1] [2]. In the multi-band SS approach, the effect of $x_E(n)$ is addressed through Cepstral Mean Normalization (CMN), which can be handled by the recognizer as it falls within the frame. Thus, only $x_L(n)$ is removed through the multi-band SS as its effect falls outside the frame in which the recognizer operates. The power spectra of $x_E(n)$ can be obtained through the multi-band SS,

$$|X_E(f,\tau)| = \begin{cases} |X(f,\tau)|^2 - \delta_k |X_L(f,\tau)|^2 \\ \quad \text{if } |X(f,\tau)|^2 - \delta_k |X_L(f,\tau)|^2 > 0 \\ \beta |X_L(f,\tau)|^2 \quad otherwise \end{cases} \quad (3)$$

for $f \in B_k$ where B_k is the corresponding band, with β the flooring coefficient. $|X(f,\tau)|^2$ and $|X_L(f,\tau)|^2$ are the power spectra of the reverberant signal and its late reflection, respectively. The values of δ coefficients are derived through an offline training which minimizes the error of the estimate $|X_L(f,\tau)|$ under the MMSE criterion. Details in the choice of the number of bands, the values of δ coefficients (through offline training), and the effective identification of the late components of the impulse response $h_L(n)$ are discussed in [1] [2].

3 Optimization of Dereverberation Parameters for Acoustic Modeling

The conventional approach adopts MMSE in deriving the coefficients used in dereverberation. The derived coefficients are used to process the reverberant signal, and then the acoustic model is trained using the enhanced data. We present two methods that optimize the dereverberation parameters jointly with acoustic modeling. This principle is also applied during actual recognition which will be discussed in Section 4. The two methods are explained as follows:

3.1 Batch Optimization Method

The proposed optimization of the multi-band SS is shown in Fig. 3. We opt to optimize each band sequentially starting from the first band $m = 1$ to $m = M$. The band coefficient to be optimized is allowed to change

Table 1: System specifications

Sampling frequency	16 kHz
Window Frame length	25 ms
Window Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCCs, 12-order Δ MFCCs 1-order Δ E
HMM	8000 Gaussian pdfs
Training database	Male and Female Adult by JNAS
Test data	Male and Female Adult by JNAS

Table 2: Basic Recognition Results

Methods	200 msec	600 msec
(A) No processing	68.6 %	44.0%
(B) Conventional: MMSE	**80.1** %	**62.3%**
(C) Batch (training only)	81.3 %	64.3%
(D) Incremental (training only)	82.4 %	65.4%
(E) Batch (training/decoding)	83.1 %	66.1%
(F) Incremental (training/decoding)	**84.5** %	**67.5%**

within a close neighborhood $n\Delta$ where $n = 1...N$ and $\Delta = 0.02$. The reverberant observation data $\boldsymbol{x}$ is dereverberated using the multi-band SS. The rest of the bands are fixed to the MMSE-based estimates except for the band to be optimized. Thus, if the band to be optimized is band $m = 1$, we generate a set of coefficients $\delta(1,n) = [\,\delta(1)_{MMSE} + n\,\Delta,\ \delta(2)_{MMSE},\ \delta(m)_{MMSE}$ $,...,\ \delta(M)_{MMSE}]$, and execute SS using the generated coefficients. The resulting data $x_E(\delta(1,n))$ are evaluated using the HMM-based acoustic model which is trained with data processed with MMSE-based SS parameters, denoted as $\lambda = \lambda_{MMSE}$. A Likelihood score is computed for each of the data processed with different SS conditions. Based on this result, $\delta(m)_{opt}$ that has the corresponding highest likelihood score is selected. The whole process from SS to likelihood evaluation is applied to all M bands independently. After all of the bands are optimized, the set of optimal SS coefficients $[\delta(1)_{opt}, ..., \delta(M)_{opt}]$ is used to process the reverberant data and proceed to acoustic model training. The resulting acoustic model will be used in the actual recognition.

3.2 Incremental Optimization Method

We extend the above *batch optimization method*. The additional process introduced is shown in dashed lines in Fig 3. Right after the optimal coefficient of band 1 is found, the acoustic model is re-estimated using the updated SS parameters. The newly re-estimated model λ_1 is then used in the likelihood evaluation block for band 2, and this process is iterated until $\delta(M)_{opt}$ is found for the Mth band. This approach, referred to as *incremental optimization method*, has the same principle with the *batch method*, except for the incremental updates of the HMM parameter λ in every band. In the *batch method*, we fixed $\lambda = \lambda_{MMSE}$ all throughout the bands. The incremental re-estimation allows us to treat each band interdependently in a sequential manner as opposed to the *batch optimization method* where each band is treated independently.

4 Optimal Parameter Selection During Decoding

Further optimization is implemented during actual recognition. Using the acoustic model processed with the optimal multi-band SS parameters in section 3, we evaluate a likelihood given a dereverberated test utterance. The reverberant test data are processed in the same manner as the optimization of the bands in the acoustic training phase, producing a set of processed utterances. These utterances are then evaluated with the acoustic model. The corresponding multi-band coefficient that gives the highest likelihood is selected for each band which is similar to that shown in Fig 3, and used for the final recognition. Since the dereverberation based on the multi-band SS depends on the room impulse response measurement, it is possible that the initial condition of the room impulse response used in training the model is not maintained in the actual recognition. Thus, the additional optimization during decoding is beneficial to the system in minimizing the mismatch between the actual test data and the acoustic model.

5 Experimental Evaluation

For evaluation of the proposed method, we used the training database from Japanese Newspaper Article Sentence (JNAS) corpus. The test set is composed of 200 utterances taken outside of the training database. Recognition experiments are carried out on the Japanese dictation task with 20K-word vocabulary. System specifi-

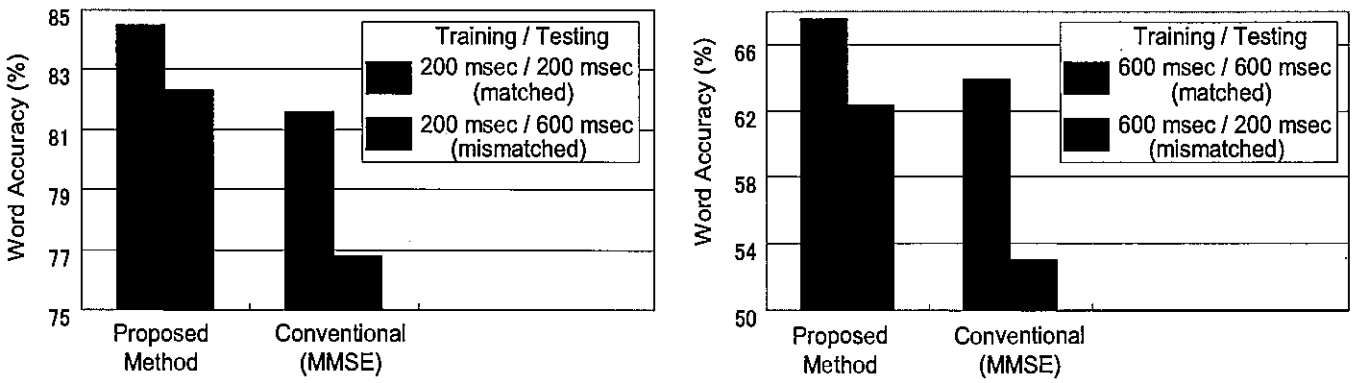


Figure 4: Test for robustness

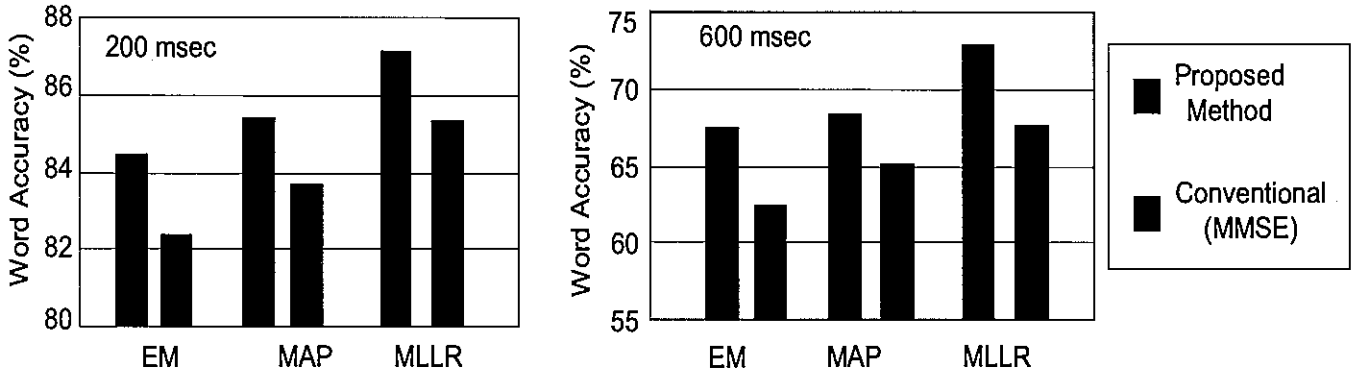


Figure 5: Performance when used in adaptation

cation is summarized in Table 1. The language model is a standard word trigram model. We experimented using two reverberant conditions: 200 msec and 600 msec. Reverberant data were made by convolving the clean database with the measured room impulse response [11]. The measured room impulse response contained flutter echo which is inherent of the actual room acoustics. In this experiment we use total number of bands $M = 5$ which is consistent to that of the former work [1][2].

5.1 Recognition Performance

Table 2 shows the basic recognition performance (word accuracy) of the proposed method in 200 msec and 600 msec reverberant conditions. (A) is the performance for reverberant test data (without dereverberation) using a clean acoustic model. (B) is for the conventional MMSE-based approach when both the test and training data are dereverberated with the conventional MMSE-based SS. (C) and (D) are the results of the proposed optimization for the batch and incremental methods, respectively. It is confirmed that the proposed front-end dereverberation optimization considering acoustic likelihood is more effective than the conventional MMSE-based method. And the incremental model update performs better than the batch training. In (E) and (F), we show that the performance of the system is further improved when optimization is also applied in the decoding process. Thus, optimizing dereverberation in both the acoustic modeling phase and decoding phase result in a synergetic effect in improving recognition accuracy. As a whole, we have achieved a relative 5% improvement over the baseline MMSE-based method.

5.2 Robustness of the Proposed Method

We also performed experiments regarding the robustness of the proposed approach. In real environment condition, it is possible that room impulse response may have considerably changed due to the additional presence/absence of physical fixtures inside the room which were absent during the measurement causing a mismatch between the acoustic model and the test data. By using different impulse responses in creating the reverberant test data and the training data, we simulate a mismatch of the reverberant condition and investigate the robustness of the proposed method as shown in Fig. 4. It is apparent that the change in the recognition performance from (matched) to (mismatched) is much smaller under the proposed method than in the conventional approach using MMSE criterion. We note that unlike the conventional method, the proposed approach is capable of optimizing the dereverberation parameters during the actual recognition which can further minimize mismatch.

5.3 Evaluation with MAP and MLLR

Then, we extend the proposed optimization technique to the adaptation scheme like MAP and MLLR. In this

case, we execute an iterative MAP and MLLR, and in each iteration we optimize the dereverberation parameters together with the 50 adaptation utterances. Recognition results shown in Figure 5 demonstrates that the proposed approach is effective in conjunction with adaptation, especially with MLLR, and the advantage over the conventional method is maintained after the adaptation.

5.4 Faster Implementation of the Proposed Optimization Technique

The proposed optimization process outlined in Fig 3 that uses HMM in evaluating the likelihood is confirmed to be effective in optimizing the dereverberation parameters. However, this process takes a lot of time and it is desirable to replicate the same performance in a shorter period of time. We try to use Gaussian Mixture Model (GMM) with 64 mixture components instead of HMM in finding the optimal parameters. A separate HMM is trained/updated only after the optimal parameters are found through GMM. This means that GMM is used for the optimization process and HMM is used for the actual speech recognition. This approach has been shown to be effective in VTLN [10].

In Fig. 6, we show the result for using both GMM and HMM in finding the optimal multi-band SS parameters. We can observe a negligible difference in word accuracy between GMM and HMM. With the GMM implementation, we reduced optimization time up to 10%. This implementation makes decoding in section 4 practical.

6 Conclusion

We have presented the front-end dereverberation technique which is optimized based on the likelihood of the speech recognizer. The proposed is applied to the acoustic model training phase and the actual decoding phase. Both effects are confirmed, realizing significantly better performance than the conventional MMSE-based method which optimizes the parameters independent of speech recognition. We have also presented a method of speeding up the optimization process through the use of GMM which renders the decoding to be fast. In our future works, we will expand the current approach to an unknown room impulse response, where we can replace the room acoustics dependency with recognizer-based optimization in enhancing the reverberant speech signal for robust speech recognition.

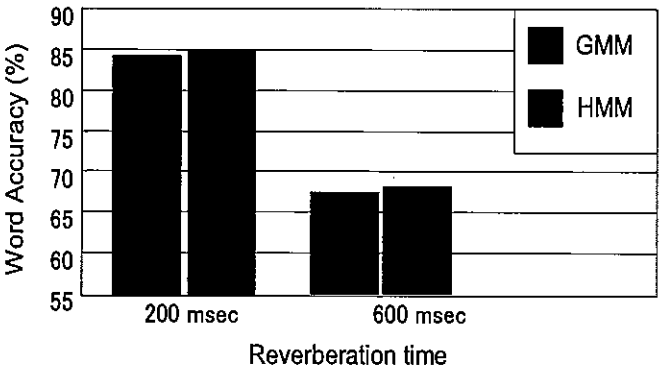


Figure 6: Performance comparison between GMM and HMM in optimizing the multi-band coefficients

References

[1] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Distant-talking Robust Speech Recognition Using Late Reflection Components of Room Impulse Response" *In Proceeding of International Conference on Acoustics Speech and Signal Processing*, 2008

[2] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Fast Dereverberation for Hands-Free Speech Recognition" *IEEE Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, 2008

[3] K. Kinoshita, T. Nakatani and M. Miyoshi, "Spectral Subtraction Steered By Multi-step Forward Linear Prediction For Single Channel Speech Dereverberation" *In Proceeding of International Conference on Acoustics Speech and Signal Processing*, 2006

[4] A. Acero and R.M. Stern, "Environmental Robustness in Automatic Speech Recognition" *In Proceeding of International Conference on Acoustics Speech and Signal Processing*, pp 849-852 1990

[5] A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition" *Kluwer Academic Publishers, Boston, MA*, 1993

[3] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Efficient Dereverberation Framework For Automatic Speech Recognition" *In Proceedings of Interspeech* , Vol 1, pp 92-95, 2005

[6] M. Seltzer, "Speech-Recognizer-Based Optimization for Microphone Array Processing" *IEEE Signal Processing Letters*, Vol. 10, No. 3, 2003

[7] M. Seltzer and R. Stern, "Subband Likelihood-Maximizing Beamforming for Speech Recognition in Reverberant Environments" *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 6, 2006

[8] L. Lee and R. Rose, "Speaker Normalization using Efficient Frequency Warping Procedures" *In Proceeding of International Conference on Acoustics Speech and Signal Processing*, pp 353-356, 1996

[9] D.Pye and P.C.Woodland, "Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition" *In Proceeding of International Conference on Acoustics Speech and Signal Processing*, pp 1047-1050, 1997

[10] L. Welling, H. Ney, and S. Kanthak, "Speaker Adaptive Modeling by Vocal Tract Normalization" *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 10, No. 6, 2002

[11] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses" *Journal of Acoustical Society of America. Vol.97(2), pp.-1119-1123, 1995*

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A802-5 (11/18)

実環境におけるMUSIC法を用いた3次元音源定位の評価

Evaluation of the MUSIC algorithm on 3D Localization of Sound Sources in Real Noisy Environments

○石井カルロス寿憲（ＡＴＲ知能ロボティクス研究所）
シャトッ・オリビエ（MIT，ＡＴＲ知能ロボティクス研究所）
石黒浩（大阪大学工学部，ＡＴＲ知能ロボティクス研究所）
萩田紀博（ＡＴＲ知能ロボティクス研究所）

* Carlos Toshinori ISHI, Olivier CHATOT, Hiroshi ISHIGURO, Norihiro HAGITA (Intelligent Robotics and Communication Labs., ATR)

carlos@atr.jp, ochatot@mit.edu, ishiguro@ams.eng.osaka-u.ac.jp, hagita@atr.jp

Abstract - With the goal of improving human-robot speech communication, a 3D sound source localization based on the MUSIC algorithm was implemented and evaluated in our humanoid robot embedded in real noisy environments. The effects of the determination of the correct number of sources in the MUSIC algorithm are evaluated, and the eigenvalue profiles for each number of sources are analyzed for recordings in different environments. Based on the analysis results, a classifier was proposed for automatic determination of the number of sources using eigenvalues obtained from different frequency ranges. Evaluation results showed that a combination of two sets of eigenvalues calculated by averaging the eigenvalues of the frequency bins in two separate ranges resulted in the best performances.

1　はじめに

人とロボットとの音声コミュニケーションにおいて、ロボットに取り付けたマイクロホンは通常離れた位置（１ｍ以上）にあり、例えば電話音声のようにマイクと口との距離が数センチの場合と比べて、信号と雑音の比（SNR）は低くなる。このため、傍にいる他人の声や環境の雑音が妨害音となり、ロボットによる目的音声の認識が難しくなる。従って、ロボットへの応用として、音源定位や音源分離は重要となる。

音源定位に関しては過去にさまざまな研究がされている[1]-[11]。しかしながら、その大半ではシミュレーション・データや、ラボ・データのみ使用され、ロボットが動作する実環境のデータを評価するものは少ない。また、３次元の音源定位を評価する研究も少ない[8]-[9]。発話相手の顔を見ながら話す・聞くことも人間とロボットの対話インタラクションを改善するための重要なビヘービアであり、そのためには３次元の音源定位も重要となる。

以上の実状を踏まえ、本研究では、我々の研究室の人型コミュニケーションロボット「ロボビー」に３次元の音源定位を実装し、実環境の雑音環境で評価を行った。本研究では、分解能が高いMUSIC法(MUltiple SIgnal Classification)と呼ばれる有名な音源定位の手法を扱った[1]-[3]。環境の変化によるMUSIC法の問題を分析し、音源定位のロバスト化を改善するための手法を提案した。

本稿は以下のように構成される。次ぐ第２節ではハードウェアと収録データを記述する。第３節では、提案手法を説明し、第４節では分析と評価結果を示す。第５節で結論を述べる。

2　ハードウェアおよび収録データ

2.1 マイクロホンアレイ

14個のマイクロホンによるアレイを、図１に示すように、ロボビーの胸部にフィットするよう作成した。さまざまな３次元のアレイ構造におけるMUSIC法のシミュレーションを試した結果、MUSIC出力のサイドローブが最も少なく、低域および高域の周波数帯域で良い応答を示した図１に示すアレイを採用した。ただし、これ以降の分析では、空間的分解能が低い１ｋHz以下の帯域と、空間的aliasingが起こり得る６ｋHz以上の帯域を除外した。

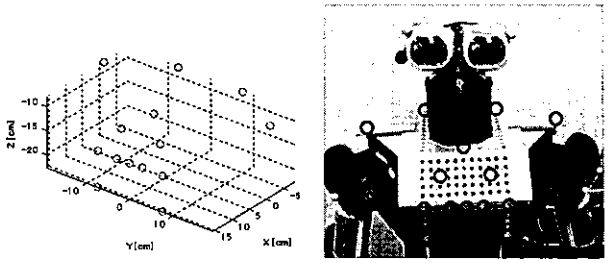


Fig. 1. (a) The geometry of the 14-element microphone array. (b) Robovie wearing the microphone array.

マイクロホンアレイの信号をキャプチャするため、Tokyo Electron Device Limited の

TD-BD-16ADUSB という１６チャンネルのA/D変換機を用いた。マイクロホンは、Sonyの無指向性のコンデンサーマイク ECM-C10 を用いた。オーディオ信号は、音声認識で一般的に使用される 16 kHz/16 bit でキャプチャした。

2.2 実験のセットアップ

マイクロホンアレイをロボビーの胸部にフィットさせた。ロボットの内部雑音も考慮させるため、ロボットの電源は入れた状態にした。音源となる話者らはロボットの周りのさまざまな方位に配置し、ロボットに向かって自然に発話するよう指示した。各音源のレファレンスとなる信号を求めるため、各話者には追加のピンマイクロホンを持たせた。これらの追加のマイクロホンから得られた信号を本稿で「音源信号」と呼ぶ。これらの音源信号は、分析と評価に用いるためであり、最終的な実装には不要である。

2.3 データ収集および環境の条件

マイクロホンアレイによるデータ収録を２つの異なった環境で行った。一つ目はオフィス環境(OFC)で、室内のエアコンとロボットの内部雑音が主な雑音源となる。二つ目の環境は、現在ロボビーの実証実験が行われている「ユニバーサルセティウォーク大阪」という野外のショッピングモールの通路(UCW)である[10]。UCWでの主な雑音源は、天井に設置されているスピーカーから流れてくるポップ・ロックミュージックとなる。通路内のさまざまな位置およびさまざまな向きで収録を行った。

本稿では、その中の４つ収録の結果を示す。１つ目はオフィス環境（OFC）で、残りの３つはショッピングモールの環境（UCW1~3）である。表１にそれぞれの収録での音源に関する詳細を示す。

TABLE I
APPROXIMATE POSITIONS OF THE SOURCES FOR EACH RECORDING.

Recording	Approximate positions of the sources
OFC	Sources at -50, -20, +20 and +50 degrees relative to the robot.
UCW1	Sources at +30 degrees and -30 degrees. Robot located away from the ceil loudspeakers (music sources).
UCW2	Sources at +25 and -25 degrees. Robot located right under a loudspeaker (music source).
UCW3	Source 1: moves from 50 to -90, and back. Source 2: moves from -20 to 0 after 60 time blocks (when source 1 crosses it). Music: at ahout -45 degrees, with the volume (casually) decreased in the second half of the recording.

OFCでは、４つの音源（男性話者）が存在する。まず、各話者が10秒間程度ロボットに話しかけた（他の話者は静かにした）。最後の15秒間に全4話者が同時に話しかけた。この収録では、2人の話者が16チャンネルのA/D変換機の余りの2個のチャンネルに接続したマイクロホンを使用し、残りの2人の話者は別のオーディオキャプチャ装置に(M-audio USB audio）に接続したマイクロホンを使用した。収録の初めに一度手を叩くことにより、残りの２話者の信号を手動でアレイの信号と同期させた。

UCW の収録では、ターゲット音源が２つ（男性話者２名）が存在する。UCW1 と UCW2 では、各話者が 10 秒程度ロボットに話しかけ、その後同時に発話する。UCW3 では、一人の話者がロボットの前を移動しながら発話している。収録中、背景の音楽のボリュームが偶然に下げられた。

2.4 音源信号のパワーによる（レファレンス）音源数の推定

音源数(NOS)は、MUSIC 法が要するパラメータである。次節で紹介する音源数の推定のための分析と評価には、レファレンスとなる音源数を音源信号のパワーから推定する。認識の際に Classifier より得られる NOS と区別するため、パワーより得られるレファレンスの音源数を PNOS と呼ぶ。

各音源のパワーを求める前に、音源信号間でクロス・チャンネル・スペクトル・バイナリー・マスキングを行った。（具体的には、２つの信号を周波数領域に変換し、個々の周波数成分を比較し、強いものは残し、弱いものにはゼロを割り当て、時間領域に戻す処理である。）この処理により、チャンネル間の音漏れを抑え、より信頼性の高いレファレンス信号が求められる。更には、音源信号から環境音を除外するため、アレイの中央に位置するマイクロホンの信号を用いてすべての音源信号にバイナリー・マスキングを行った。

音源信号のパワーは25msのフレーム毎に求めた。パワー軌道に閾値を設定することで音源信号がアクティブであるかを判断した。アクティブである音源の数がそのフレームの音源数（PNOS）となる。

3 提案手法

3.1 MUSIC応答の推定

各音源の方位角のみならず、仰角も推定するため、MUSICアルゴリズム[1]（付録１参照）の3次元版を実装した。方位角と仰角のセットをこれ以降音源方位（DOA）と呼ぶ。本研究では、音源の距離推定はせず、音源方位（方位角、仰角）のみを検出することとした。これにより処理時間を大幅に減少させることができる。

方位角、仰角ともに、5度の間隔からなる網目を使用し、仰角が-30度以下の方位は計算量を減少するため削除した。座標の原点をロボビーの頭部の回転軸の交点となるよう設定した。これにより、音源定位の出力をそのままロボットの頭部を制御するのに用いられる。

[1]では、相関行列が１秒間のブロック毎に平均され、音源定位の結果も少なくとも1秒後に求められる。本研究では、出来るだけ実時間処理に近づけるため、ブロック長を200 msとした。

MUSIC法では、MUSIC応答を求める際に、どの固有ベクトルを使うかを決定するため、音源数(NOS)を予め設定する必要がある。NOSの推定およびMUSIC応答の推定は、ブロック毎に行われる。

提案手法では、音源数 NOS を相関行列の固有値をパラメータとした classifier により推定する。classifier の学習においては、2.4 節で紹介した音源のパワーから求められたレファレンスとなる音源数（PNOS）を使用する。ただし、PNOS は 25ms のフレーム毎に求められるため、平均と四捨五入により、200ms のブロックに変換した。固有値の分析および classifier の評価における詳細は 4 節で述べる。

3.2 MUSIC応答による音源方位の検出

各時間ブロックでMUSIC応答が得られると、局所的なピークを検索することにより、DOAを求める。ただし、２つの音源の方位が近い場合には、局所的なピークが１つしか存在しないことがあり得る。この問題を解決するため、以下のような方法を提案した。まず、最大の局所的ピークを検出し、2次元Gaussian（方位角と仰角）をMUSIC応答から差し引く。このGaussianは、１つの音源が存在する時の標準偏差を持ち、検出されたピークの振幅を持つものとする。この作業を音源の数だけ繰り返す。

3.3 音源方位のフィルタリングとトラッキング

音源数が過大推定された場合、誤ったDOAの挿入が起きる。提案手法では、グルーピングによるフィルタリングのアルゴリズムを用いることで、孤立したDOAを削除する。本アルゴリズムは、過去１０ブロック（2秒に相当）に検出されたDOAより、現在のDOA候補をグループ化するか否かを判断する。以下の条件にあてはまる場合、グループ化を行う。

1) 前のDOAは、現在のDOAを先端とした「円錐」の内部にある。円錐の底面は方位角が±30度、仰角が±7度に設定した。これらの値は、人は縦方向（仰角の変化）よりも、横方向（方位角の変化）に移動する確率が高いことに基づき、ヒューリステックに設定した。

2) 現在のDOAと前のDOAが属するグループの傾向線との距離がある閾値よりも小さい。

1番目の条件により、近いグループにのみグループ化可能となり、2番目の条件により、方向性が異なるグループにはグループ化しないこととなる。

図２に、1個の音源が50度から-90度の間を往復する場合(UCW3)のトラッキング結果を示す。グループ化されたDOAは濃い（青い）×点で、除外された

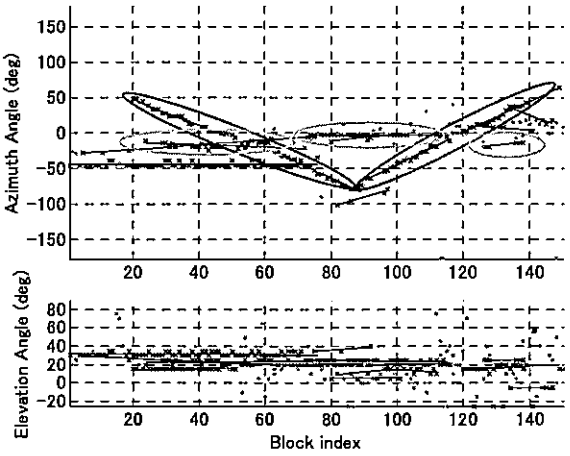


Fig. 2. Filtering and tracking results for two subjects (-20°~0°, and 50°~-80°~50°) and background music (at -45°, in the first half). Full lines are the detected groupings and ellipses show the correct source positions.

DOAは薄い（緑の）点で表示されている。直線はグループを示す。楕円で囲まれたものがターゲットとなる音源である。いずれの音源もうまくトラッキング出来たことが分かる。更に、収録の前半で指向的音源となっていた環境音楽も方位角-45度で正しく検出されている。

方位角の図では、連続しているような薄い（緑の）点が観えるが、これらの点の仰角が連続していないことがグループ化されなかった理由である。

図２では、ターゲットの音源ではないグループも検出されたことが分かる。これらの挿入グループは後続処理となる音源分離で除外できることを期待する。ただし、「はい」や「え？」のような短い発話もあり得るため、このような孤立した DOA を処理することも重要となる。

4 分析および評価実験

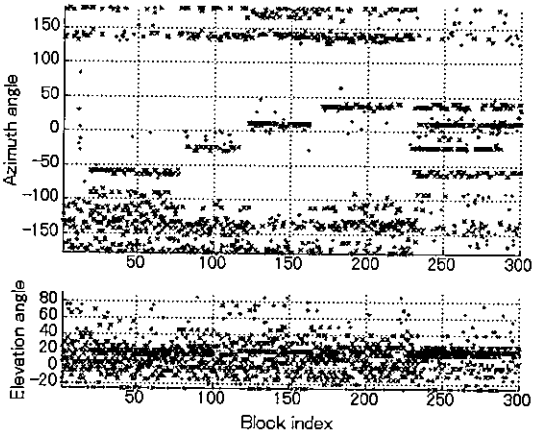
4.1 音源方位の検出における音源数の影響

結論からいうと、音源数（NOS）を適切に推定することにより、MUSIC応答から得られる音源方位の推定の後処理が容易となる。図３に、オフィスでの収録（OFC）に対し、NOSの異なった条件により、検出された音源方位（DOA）を示す。濃い（青い）×点はDOA filtering/tracking後に検出されたDOAである。図３(a)では、すべてのブロックにおいてNOSを5に固定した場合に検出された音源方位（DOA）を表示する。本収録では同時に最大４個の音源数が存在し、すべてのブロックにおいてNOSを過大評価したことに相当する。図３(b)は、レファレンスとなるPNOSを使用した場合の結果である。

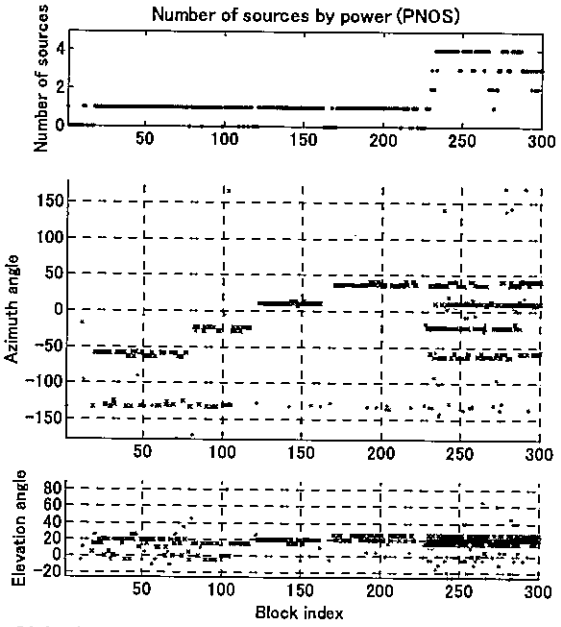
検出されたDOAを評価するため、２つの測定値を使用した。1つ目はターゲットとなるDOAの正解率を表す。この値をDOA accuracyと呼ぶ。2つ目はブロック毎に検出されたDOAの挿入誤り数の平均値で

ある。この値をANFP（Average Number of False Positives per block）と呼ぶ。

表IIに、DOA accuracy およびANFPの測定値をNOSの２つに条件において、DOA filtering/trackingを行う前後の値を表示する。NOSを固定した場合、PNOSよりも少し高い正解率（DOA accuracy）を示しているが、挿入誤り（ANFP）がかなり高い値を示している。より正確なNOSを使用することにより、ANFPを減少できることがわかる。



(a) Setting a fixed number of sources to 5, at all times.



(b) Setting number of sources by power of the sources signals (PNOS).

Fig.3. Detected DOA over time for the trial in the office environment (OFC, 4 subjects between -60° and 50° azimuth).

TABLE II
EFFECTS OF NOS IN DOA ESTIMATION, FOR OFC RECORDING.

NOS	DOA filtering/tracking	DOA accuracy (%)	ANFP
Fixed NOS = 5	before	85.0	4.60
	after	80.6	3.15
PNOS	before	80.6	0.32
	after	73.9	0.12

4.2 異なる環境における固有値の分析

適切なMUSIC応答を求めるためには、音源数を推定する必要がある。MUSIC法の過程でブロック毎に得られる、観測信号の相関行列の固有値が、音源数に関連することは知られる[1]。

本節では、固有値が、環境の変化によりどのように影響されるかを調べるため、PNOS毎に固有値を整理した。各周波数点（frequency bin）に固有値のセットが得られるが、本手法では、ブロック毎に１つの代表的な固有値のセットを求めるため、特定の周波数帯域で平均化した固有値をそのブロックの固有値プロファイルとして扱う。

図４に、３つの異なった環境（OFC、UCW1、UCW2）において、PNOS毎に整理したブロック毎の固有値のプロファイルを表示する。音源数（NOS）が、固有値プロファイルの全般的なoffsetおよび形状に関連していることが観察される。理想的な固有値プロファイルの形状としては、指向的な音源の数Nに対応する最初のN個の固有値が強いパワーを示し、無指向的な音源に対応する残りのM-N個の固有値が小さいパワーを示す。しかしながら、図4に表示している実際の形状では、指向的音源の成分と無指向的音源の成分の境目が不明確であり、無指向成分もフラットではなく、緩やかな傾きを示している。

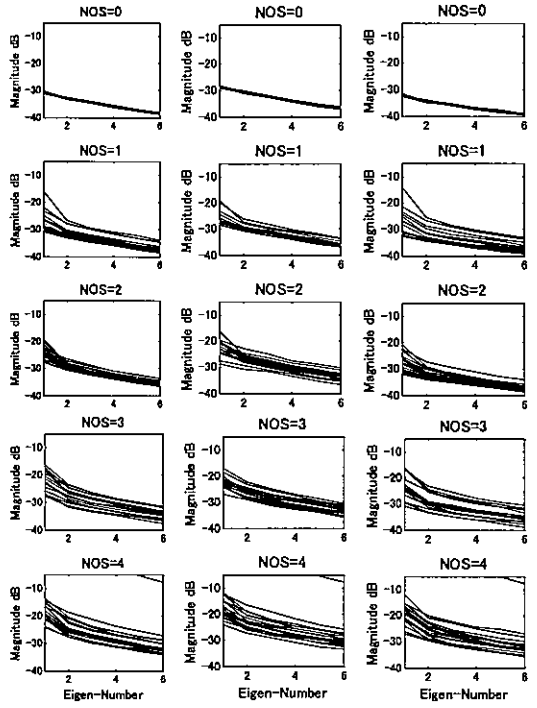
また、異なったPNOS間でも固有値プロファイルが一部重複することも観察される。例えば、OFC (図4(a))では、PNOS＝１とPNOS＝２のプロファイルが大幅に重複していることが観察される。従って、classifierによる厳密な音源数の推定は困難であることが予想される。

更には、環境の変化による、固有値プロファイルの形状への影響も強いことが観られる。幅にも傾きにも違いが観られる。例えば、OFCとUCW1のPNOS=0の固有値プロファイル（図４(a),(b)のNOS=0）を比較すると、その違いは明らかである。UCW1では、背景の音楽があるため、OFCよりも値が大きく、ばらつきも大きくなっている。これは環境の変化をclassifierに考慮する必要があることを示している。

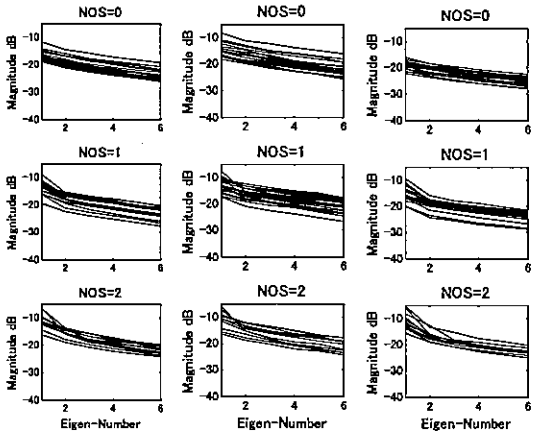
また、環境音楽の音源に近づくことによる固有値プロファイルへの影響も観られる。UCW2のPNOS=0のプロファイルの形状は、UCW1のPNOS=1のものと類似している。これは、ロボットが環境音楽の音源に近づく場合は、環境音楽を新たな指向的音源となり、離れている場合は、無指向的音源となることが反映されている。環境音楽が指向的音源の場合、その方位を求めることが出来、後続処理となるターゲット音声の音源分離にも役立つ。

最後に、異なった周波数帯域で平均化した固有値のプロファイルを分析した。図４の3列に、それぞれ 1～6 kHz (AVG1_6)、1～3 kHz (AVG1_3) および

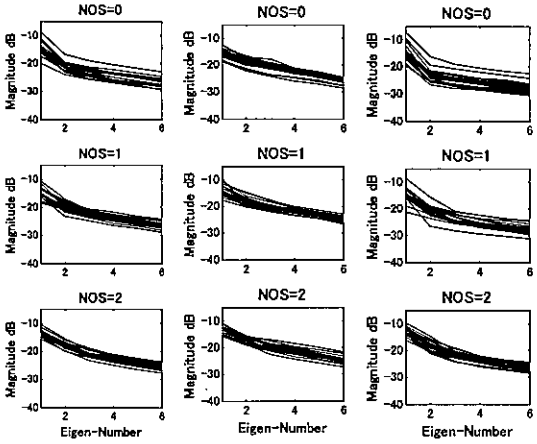
3 ~ 6 kHz (AVG3_6) の3つの異なった周波数帯域の周波数ビンで平均化した固有値のプロファイルを



(a) OFC (PNOS = 0 ~ 4).



(b) UCW1 (PNOS = 0 ~ 2).



(c) UCW2 (PNOS = 0 ~ 2).

Fig. 4 Eigenvalue profiles for different environments, arranged by PNOS. The three columns in the figure are AVG1_6, AVG1_3, and AVG3_6, respectively. The eigenvalues larger than 6 were removed from the plots, because they didn't show significant change in their shapes.

示す。NOS>0のプロファイルで図4の3列を比較すると、AVG3_6（右の列）では、第1と第6の固有値の差がより大きいことが分かる。この結果より、AVG3_6の方が、帯域幅が広域であるAVG1_6よりも高い識別性を持つと考えられる。しかし、/u/や/o/のように3kHz以上の成分が弱い音声区間では、AVG3_6では検出されない恐れがある。これらの結果より、分割した周波数帯域から求められる固有値の２セットをclassifierに用いる方法を提案した。

4.3 固有値による音源数の推定

分類アルゴリズムとして、kNN (k-Nearest Neighbors)アルゴリズムを選択した。kNNは計算量も少なく、非線形にも対応できるためである。（線形回帰分析も試みたが、kNNの方がよい分類結果を示したため、結果を省略する。）

観測信号の相関行列から得られた固有値を分類アルゴリズムの入力パラメータとして用いる。さまざまな周波数帯域の周波数ビンを通して求めた固有値の平均値セット(AVG)もしくは最大値セット(MAX)を入力として、さまざまなclassifierを学習・評価した。周波数帯域を２つに分割して得られた固有値の２セットも評価した。

kNN classifierの性能を、さまざまなk（nearest neighborの数）に対し、10-fold cross-validationにより評価した。図５に、さまざまなclassifierにおいて、推定したNOSとレファレンスのPNOSがマッチした度合い(NOS accuracy)を表示する。図５は、最もNOS accuracyが高かったk=6の場合の結果である。

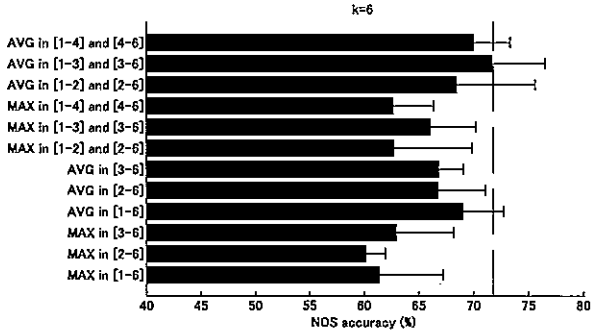


Fig. 5. Accuracy (mean and standard deviations) for estimation of number of sources by a kNN classifier with k=6, for several types of eigenvalue sets (AVG or MAX, in several frequency ranges [in kHz]).

図５より、まず固有値の最大値(MAX)よりも、平均値(AVG)の方が高い性能を示すことが分かる。最も性能が高かったのは、AVG in [1-3] and [3-6]であり、周波数帯域を分割した固有値の２セットを使用することが効果的であることを示している。

4.4 音源方位の検出におけるFFTの点数の影響

音源方位（DOA）の検出におけるFFTの点数（NFFT）の影響を調べた。図６に、NFFTのさまざまな値 (32, 64, 128, 256, 512)に対する、DOAの性能

を表示する。NFFT＝128が音源方位の推定において最も高い性能を示した。しかし、より小さいNFFTでも著しい性能の劣化は観られなかった。より小さいNFFTを使うことで計算量は大幅に減少できるため、本研究では、NFFT=64を採用した。

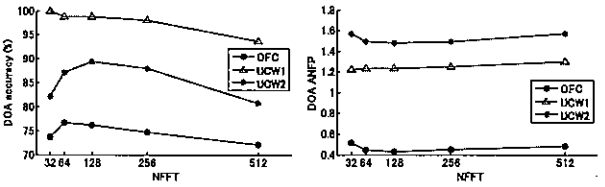


Fig. 6. Accuracy of one block as a function of different values for the number of FFT points, for recordings OFC, UCW1 and UCW2 .

5　おわりに

MUSIC 法に基づいた 3 次元音源定位を実装し、人型ロボットの実環境における評価を行った。音源数の MUSIC 法への影響を異なった環境で分析した。音源数の推定誤りは、音源方位の挿入誤りを起こす可能性があり、後続の音源分離を難しくすることとなる。環境の変化における固有値のプロファイルへの影響をさまざまな音源数において分析した。

異なった周波数帯域の周波数ビンから得られた固有値のプロファイルを入力パラメータとした kNN classifier を構築し、音源数を推定する方法を提案した。評価実験の結果、2 つの周波数帯域(1~3 kHz と 3~6 kHz) で求めた平均固有値のプロファイルで最も高い性能が得られた。

今後の課題として、音源分離を実装・分析する。また、本研究の音源方位を基に、人型ロボットの頭部を制御することを試みる。

付録：MUSIC法

M個のマイク入力のフーリエ変換$X_m(k,t)$は、式(1)のようにモデル化される。

$$\mathbf{x}(k,t)=[X_1(k,t),...,X_M(k,t)]^{\mathrm{T}}=\mathbf{A}_k\,\mathbf{s}(k,t)+\mathbf{n}(k,t) \quad (1)$$

ベクトル$\mathbf{s}(k,t)$はN個の音源のスペクトル$S_n(k,t)$から成る：$\mathbf{s}(k,t)=[S_1(k,t),\ ...,\ S_N(k,t)]^{\mathrm{T}}$ 。kとt はそれぞれ周波数と時間フレームのインデックスを示す。ベクトル$\mathbf{n}(k,t)$は背景雑音を示す。行列 $\mathbf{A}_k$ は変換関数行列であり、(m,n)要素はn番目の音源からm番目のマイクロホンへの直接パスの変換関数である。$\mathbf{A}_k$ のn列目のベクトルをn番目の音源の位置ベクトル（steering vector）と呼ぶ。

まず、式(2)で定義される空間相関行列$\mathbf{R}_k$を求め、式(3)に示す$\mathbf{R}_k$の固有値分解により、固有値の対角行列$\mathbf{\Lambda}_k$および固有ベクトルから成る$\mathbf{E}_k$が求められる。

$$\mathbf{R}_k=E[\mathbf{x}(k,t)\mathbf{x}^H(k,t)] \quad (2)$$

$$\mathbf{R}_k=\mathbf{E}_k\mathbf{\Lambda}_k\mathbf{E}_k^{-1} \quad (3)$$

固有ベクトルは $\mathbf{E}_k=[\mathbf{E}_k^{s}\mid\mathbf{E}_k^{n}]$ のように分割出来、$\mathbf{E}_k^{s}$と$\mathbf{E}_k^{n}$はそれぞれ支配的なN個の固有値に対応する固有ベクトルと、それ以外の固有ベクトルである。

MUSIC空間スペクトルは式(4)と(5)で求める。rは距離、θとφはそれぞれ方位角と仰角を示す。式(5)は、スキャンされる点 (r,θ,φ) における正規化した位置ベクトルである。

$$P(r,\theta,\varphi,k)=\frac{1}{|\,\tilde{\mathbf{a}}_k^H(r,\theta,\varphi)\mathbf{E}_k^n\,|^2} \quad (4)$$

$$\tilde{\mathbf{a}}_k(r,\theta,\varphi)=\frac{\mathbf{a}_k(r,\theta,\varphi)}{\|\mathbf{a}_k(r,\theta,\varphi)\|} \quad (5)$$

空間スペクトル（本稿ではMUSIC応答と呼ぶ）は、MUSIC空間スペクトルを式(6)のように平均化したものである。

$$\overline{P}(r,\theta,\varphi)=\frac{1}{K}\sum_{k=k_L}^{k_H}P(r,\theta,\varphi,k) \quad (6)$$

k_Lとk_Hは、周波数帯域の下位と上位の境界のインデックスであり、$K = k_H - k_L + 1$。音源の方位は、MUSIC応答のN個のピークから求められる。

謝辞

本研究は部分的に総務省および NEDO の研究委託により実施したものである。

参考文献

1) F. Asano, M. Goto, K. Itou, and H. Asoh, "Real-time sound source localization and separation system and its application on automatic speech recognition," in *Eurospeech 2001*, Aalborg, Denmark, 2001, pp. 1013–1016.
2) K. Nakadai, H. Nakajima, M. Murase, H.G. Okuno, Y. Hasegawa and H. Tsujino, "Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays," in *Proc. of IROS 2006*, Beijing, China, 2006, pp. 852–859.
3) S. Argentieri and P. Danès, "Broadband variations of the MUSIC high-resolution method for sound source localization in Robotics," in *Proc. of IROS 2007*, San Diego, CA, USA, 2007, pp. 2009–2014.
4) M. Heckmann, T. Rodermann, F. Joublin, C. Goerick, B. Schölling, "Auditory inspired binaural robust sound source localization in echoic and noisy environments," in *Proc. of IROS 2006*, Beijing, China, 2006, pp.368–373.
5) T. Rodemann, M. Heckmann, F. Joublin, C. Goerick, B. Schölling, "Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping," in *Proc. of IROS 2006*, Beijing, China, 2006, pp.860–865.
6) J. C. Murray, S. Wermter, H. R. Erwin, "Bioinspired auditory sound localization for improving the signal to noise ratio of socially interactive robots," in *Proc. of IROS 2006*, Beijing, China, 2006, pp. 1206–1211.
7) Y. Sasaki, S. Kagami, H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *Proc. of IROS 2006*, Beijing, China, 2006, pp. 380–385.
8) J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," *IEEE ICASSP 2006*, Toulouse, France, pp. IV 841–844.
9) B. Rudzyn, W. Kadous, C. Sammut, "Real time robot audition system incorporating both 3D sound source localization and voice characterization," Procs.of *ICRA 2007*, Roma, Italy, 2007, pp. 4733–4738.
10) T. Kanda, D. F. Glas, M. Shiomi, H. Ishiguro, and N. Hagita, "Who will be the customer?: A social robot that anticipates people's behavior from their trajectories," *Tenth International Conference on Ubiquitous Computing (UbiComp 2008)*, 2008.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A802-6 (11/18)

BLIND SIGNAL EXTRACTION WITH MODIFIED SPECTRAL SUBTRACTION POST-FILTER FOR THE SUPPRESSION OF BACKGROUND NOISE

Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma,
Nara, 630-0192 JAPAN
{even, sawatari, shikano}@is.naist.jp

Abstract

In this paper, we propose a new architecture for enhancing the speech in a hands-free human/machine communication scenario. First we apply frequency domain blind signal extraction (FD-BSE) to estimate the contributions of the noise and of the speech at the microphone array. Then time frequency continuous masks are computed from the FD-BSE outputs for each of the channels. These continuous masks are used to modify the spectral subtraction post-filter. Finally this modified post-filter is applied channel wise to suppress the residual diffuse noise and the speech estimate is obtained by applying a beamformer to these cleaned channels. Simulation results show that the proposed architecture can achieve a comparable SNR as conventional spectral subtraction with less distortion of the speech.

1 Introduction

In order to improve the human/machine interface, implementing hands-free speech recognition is the most natural choice. But picking the user's voice at distance is not an easy task because of noise and reverberation. Microphone array techniques were used to improve the captured speech by reducing the effect of noise and reverberation ([1, 2]). In recent years, frequency domain blind signal separation (FD-BSS) has been used with success for recovering the speech by separating the observed signals in their different components (see review paper [3]). FD-BSS is in particular efficient for speech/speech separation [4]. But in the human/machine communication where the user's voice has to be extracted from a diffuse background noise, FD-BSS gives a better estimate of the diffuse background noise than of the target speech. Consequently FD-BSS has to be combined with some post-filtering techniques in order to improve the quality of the captured speech [5, 6].

In this paper, we propose a new architecture that combines a frequency domain blind signal extraction (FD-BSE) with a modified multichannel spectral subtraction in order to suppress the diffuse background noise present in the human/machine communication scenario. First FD-BSE extracts the speech and gives an estimate of the diffuse background noise at each of the microphone. These noise estimates are used to compute time frequency continuous masks (these are different from binary masks used in [7, 8, 4]). Then a modified spectral subtraction, where the noise estimates and the subtraction parameters are modulated using the computed masks information, is applied channel wise. Finally beamforming, using the FD-BSE speech estimate, is applied to the speech components.

The proposed method is compared to FD-BSE alone and FD-BSE combined with conventional spectral subtraction in order to show that it achieves a good noise reduction in term of SNR without introducing as much distortion as the conventional spectral subtraction.

2 Estimation of speech and background noise at microphone

In the hands-free interface for human/machine communication, the user is close to the machine whereas the other signals create a diffuse background noise. The propagation of sounds from their locations of emission to the microphone array is modeled by a convolutive mixture. After applying a F points short time Fourier transform (STFT) to the observed signals, the convolutive mixture is equivalent to F instantaneous mixtures in the frequency domain. At the fth frequency bin, the observed signals are

$$X(f,t) = A(f)S(f,t)$$

where the $n \times n$ complex valued matrix $A(f)$ represents the instantaneous mixture received by the n microphone array and $S(f,t) = [s_1(f,t), \ldots, s_n(f,t)]^T$ are the emitted signal components at the fth frequency bin. t denotes the frame index. Let us consider that $s_1(f,t)$ is the target speech signal and all the other components are the background noise. Then we can decompose the observed signals in target speech at microphone array

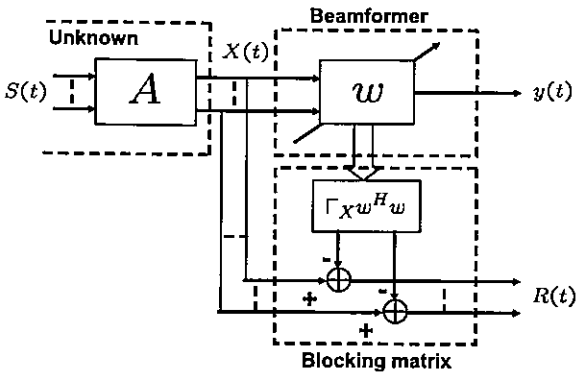


Figure 1: Blind signal extraction at frequency bin f.

and background noise at microphone array

$$\begin{aligned} X(f,t) &= A^{(1)}(f)s_1(f,t) + \sum_{i=2}^{n} A^{(i)}(f)s_n(f,t) \\ &= X_S(f,t) + X_N(f,t) \end{aligned}$$

where $A^{(i)}(f)$ denotes the ith column of $A(f)$.

The blind estimation of the speech and noise parts is possible using FD-BSS [9, 6]. Here we use the FD-BSE method proposed in [10]. Contrary to BSS, BSE estimates only one of the components of $S(f,t)$ in each frequency bin by taking

$$y(f,t) = w(f)X(f,t)$$

where $w(f)$ is a $1 \times n$ complex valued vector (see Fig.1). We call 'residuals' the contributions of all the signals other than $y(f,t)$ to the observations. The residuals are obtained by subtracting the orthogonal projection of the extracted signal from the observed signals

$$\begin{aligned} & R(f,t) = W_R(f)X(f,t), \\ \text{where} \quad & W_R(f) = I - \Gamma_X(f)w(f)^H w(f) \\ \text{with} \quad & \Gamma_X(f) = \mathcal{E}\{X(f,t)X^H(f,t)\}. \end{aligned}$$

The FD-BSE method can be seen as an adaptive beamformer and a blocking matrix as shown in Fig. 1.

In each frequency bin, the vector $w(f)$ extracting the speech component is iteratively determined using the update rule (dropping frame and frequency indexes)

$$w_{k+1} = w_k - \mu_k \mathcal{E}\{\phi(y)R^H\} W_R \qquad (1)$$

where k is the iteration index, $\mu_k > 0$ is the adaptation step and $\phi(\cdot)$ is the score function associated with the extracted component. In the frequency domain, we can assume that all the components are circular (i.e. the joint density of their modulus and phase is separable) and use the approximation $\phi(y) = \tanh(|y|)\frac{y}{|y|}$ that is appropriate for speech extraction [11]. This update rule results in an extracted signal statistically independent of the residuals.

In the human/machine scenario, the speech extraction also uses the fact that the speech distribution is spikier than that of the diffuse background noise (To measure the spikiness of the distribution, we determine the parameter of the exponential distribution fitting the normalized modulus of $y(f,t)$ and $r_i(f,t)$ [10]). When $w(f)$ is such that the speech component is extracted, the residuals are estimates of the diffuse background noise at the microphone (equivalent to the projection back of FD-BSS). The extracted speech is also projected back to the microphone array. Namely we have

$$\begin{aligned} \widehat{X_S}(f,t) &= \Gamma_X(f)w(f)^H w(f)X(f,t) \\ \widehat{X_N}(f,t) &= \left(I - \Gamma_X(f)w(f)^H w(f)\right) X(f,t). \end{aligned}$$

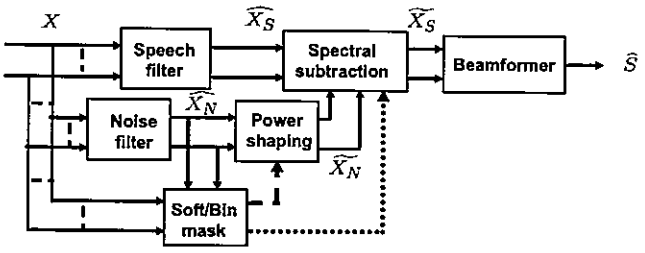


Figure 2: Proposed architecture.

3 PROPOSED ARCHITECTURE

3.1 Overview

In the proposed architecture, all the processing is performed in the frequency domain by applying a short time Fourier transform to the observed signal received by the microphone array before processing and using overlap-add method to get the time domain signal after processing.

The block diagram in Fig 2 shows the processing in the frequency domain. First FD-BSE is used to obtain the estimate of $X_S(f,t)$ and $X_N(f,t)$ denoted by $\widehat{X_S}(f,t)$ and $\widehat{X_N}(f,t)$. The noise estimate and the observation are used to determine two type of masks: Soft masks (dotted line) and binary mask (dashed line). The spectrum of the noise estimate is modified using the binary mask (the observation is also used but the arrows to the power shaping block were omitted). Then the modified spectral subtraction is performed channel wise using the shaped noise spectrum and the soft masks.

Finally, after channel wise spectral subtraction, the channels are beamformed using the vector $w(f)$ determined by the FD-BSE part.

3.2 Soft masks creation

In the human/machine communication scenario, FD-BSS or FD-BSE give a good estimate $\widehat{X_N}(f,t)$ as it is possible to cancel the speech with a spatial null [12, 6]. Then considering a frame t_i where the speech is not active $X(f,t_i) = X_N(f,t_i)$ and $X(f,t_i) \approx \widehat{X_N}(f,t_i)$. On the contrary the more the speech is active in a given frame, the more $X(f,t_i)$ and $\widehat{X_N}(f,t_i)$ differ.

Thus we propose to use the ratio of the power of $\widehat{X_N}(f,t)$ and $X(f,t)$ for a given frame as our belief in the fact that the frame is composed of noise only. Thus we define the frame soft mask as

$$P_f(t) = \frac{\sum_{f=1}^{F} |\widehat{X_N}(f,t)|^2}{\sum_{f=1}^{F} |X(f,t)|^2}.$$

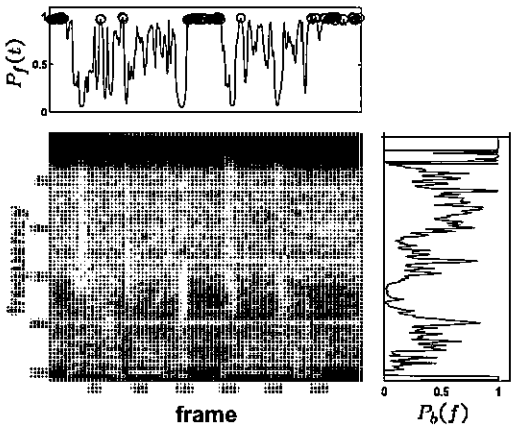


Figure 3: Observed spectrum and corresponding frame and bin soft masks.

$P_f(t_i)$ measures our belief that during the frame t_i the speech is inactive. The frame soft mask can also be seen as a measure of the frame SNR as

$$P_f(t) \approx (1 + \mathrm{SNR}(t))^{-1}$$

where $\mathrm{SNR}(t)$ is ratio of speech and noise power in the frame.

In the remainder, we also define the γ% frame binary mask obtained by selecting the γ% most probable noise frames (binary mask set to one).

Similarly, by considering the frequency bins, we can define a bin soft mask

$$P_b(f) = \frac{\sum_{t=1}^{T} |\widehat{X_N}(f,t)|^2}{\sum_{t=1}^{T} |X(f,t)|^2}$$

that measures our belief that the speech is inactive in a given frequency bin.

The frame and bin soft masks are shown along with the observed signal in Fig. 3. The circle markers on the frame soft mask indicate frames selected for the 10% frame binary mask.

3.3 Power shaping

The role of the power shaping block is to match the estimated noise and the observed signal statistics for the frames we consider as noise only. This is done by setting the mean and variance of the spectrum of $\widehat{X_N}(f,t)$, computed for the frames selected by the γ% frame binary mask, to the same values as the mean and variance of the spectrum of $X(f,t)$ for these frames.

3.4 Modified spectral subtraction

In each channel, the spectrum of the component of the power shaped noise estimate $\widetilde{X_N}(f,t)$ is subtracted from the spectrum of the component of the estimated speech $\widehat{X_S}(f,t)$

$$|\widetilde{X_S}(f,t)|^2 = \begin{cases} \text{if } |\widehat{X_S}(f,t)|^2 - H(f,t)|\widetilde{X_N}(f,t)|^2 > 0 \\ |\widehat{X_S}(f,t)|^2 - H(f,t)|\widetilde{X_N}(f,t)|^2 \\ \text{else} \\ \beta|\widetilde{X_L}(f,t)|^2 \end{cases}$$

with β the flooring coefficient. Note in particular that the subtraction parameter (referred to as α in Sect.4) of conventional spectral subtraction [13] is replaced by a mask of the noise spectrum defined by

$$H(f,t) = \delta_0 I + \delta_m P_b(f) P_f(t),$$

where δ_0 is the minimal subtraction and δ_m the additional subtraction modulated by the soft masks. Since $P_b(f)P_f(t)$ measures our belief in the absence of speech for a given time frequency value, the modified spectral subtraction only applies strong over subtraction where we believe there is no speech.

4 EXPERIMENTAL RESULTS

To demonstrate the effectiveness of the proposed approach we compared it to FD-BSE alone and FD-BSS with channel wise conventional spectral subtraction. A four microphone array (inter mic. spacing of 2.15cm) was used to record a diffuse noise (a vacuum cleaner at two meters from the array and $-40°$) and several impulse responses (at one meter from the array with angles in $[-80°, 80°]$, see Fig. 4). The room reverberation time is T60 = 200ms.

The recorded noise was mixed at different SNR with the convolution of the impulse responses and clean speech (20 signals from a database of Japanese utterances at 16kHz).

For the proposed method, three different γ% frame binary masks are considered 70%, 40% and 20% (respectively prop 1, 2 and 3 in Fig. 5). The modified subtraction parameters are $\beta = 0.003$, $\delta_0 I = 1$ and $\delta_m = 5$. The short time Fourier transform uses a 512 point hamming window with 50% overlap and pre-emphasis (a first order high pass filter $z_p = 0.97$). Speech extraction is performed by 600 iterations of the FD-BSE method with adaptation step of 0.3 divided by two every 200 iterations.

For the conventional spectral subtraction the flooring is 0.003 and the subtraction parameter is $\alpha = 2$ (mild over-subtraction) or $\alpha = 5$ (strong over-subtraction).

The proposed method being highly non linear, the SNR estimation after processing is obtained by taking

$$\mathrm{SNR} = \left(\frac{< y x_S >}{< y x_N >}\right)^2 \frac{< x_N x_N >}{< x_S x_S >},$$

where y is the output of the method and x_s and x_N are the true speech and noise at the microphone ($< \cdot >$ denotes time average).

Figure 5 shows the SNR and cepstral distortion for the speech estimate obtained with the different methods

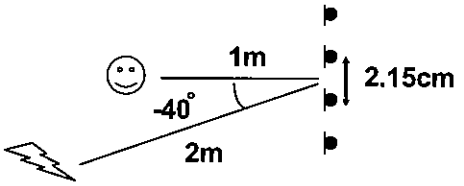


Figure 4: Room setting.

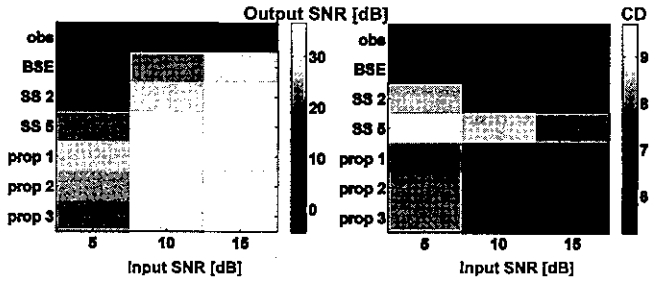


Figure 5: Averaged performance at different input SNR.

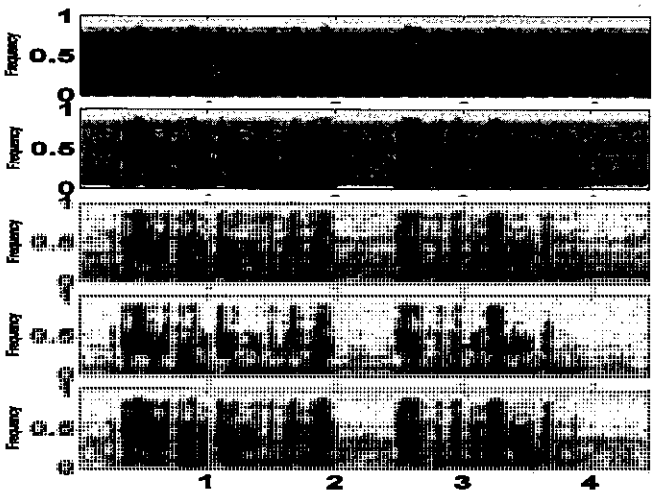


Figure 6: Spectrograms of (from top to bottom) the observation (mic. 1, 15 dB), beamformer, SS with $\alpha = 2$, SS with $\alpha = 5$ and proposed with $\gamma = 40\%$.

('obs' refers to the observation with pre-emphasis, 'BSE' to FD-BSE, 'SS 2' to conventional spectral subtraction with $\alpha = 2$, 'SS 5' to conventional spectral subtraction with $\alpha = 5$, 'prop 1' to the proposed method with $\gamma = 70\%$, 'prop 2' to the proposed method with $\gamma = 40\%$ and 'prop 3' to the proposed method with $\gamma = 20\%$). The results are averaged on all speech signals and for all the positions of the speaker. The best compromise between high SNR and low distortion is the proposed method with $\gamma = 40\%$ as FD-BSE alone introduce few distortion but does not improve significantly the SNR and conventional spectral subtraction results in higher distortion for comparable SNR (note that all signals are pre-emphasized thus input SNR and 'obs' SNR differ).

Figure 6 shows the spectrograms of the different speech estimate for an input SNR of 15 dB. We can see that the best noise reduction is obtained for 'SS 5' and the proposed method but 'SS 5' distorts the speech more than the proposed method.

5 conclusion

In this paper, considering the suppression of the diffuse background noise in the human/machine communication scenario, we proposed an architecture that achieves high SNR but introduces few distortion to the speech estimate.

References

[1] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagation*, vol. AP-30, pp. 27–34, 1982.

[2] S. Doclo et al., "Efficient frequency-domain implementation of speech distortion weighted multichannel wiener filtering for noise reduction," *EUSIPCO'04*, pp. 2007–2010, 2004.

[3] M.S. Pedersen et al., "A survey of convolutive blind source separation methods," *Springer Handbook on Speech Comm.*, 2007.

[4] Y. Mori et al., "Blind source separation combining simo-ica and simo-model-based binary masking," *ICASSP'06*, pp. 81–84, 2006.

[5] J. Kocinski, "Speech intelligibility improvement using convolutive blind source separation assisted by denoising algorithms," *Speech Communication*, vol. 50, pp. 29–37, 2008.

[6] Y. Takahashi et al., "Blind spatial subtraction array with independent component analysis for hands-free speech recognition," *IWAENC'06*, 2006.

[7] R. Lyon, "A computational model of binaural localization and separation," *ICASSP 83*, pp. 1148–1151, 1983.

[8] N. Roman et al., "Speech segregation based on sound localization," *IJCNN 01*, pp. 2861–2866, 2001.

[9] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[10] J. Even et al., "Frequency domain blind signal extraction: Application to fast estimation of diffuse background noise," *HSCMA'08, Trent, Italy*, pp. 212–215, 2008.

[11] H. Sawada et al., "Polar coordinate based nonlinear function for frequency-domain blind source separation," *IEICE Trans. Fundamentals*, vol. E86-A, no. 3, pp. 590–596, 2003.

[12] H. Saruwatari et al., "Blind source separation combining independent component analysis and beamforming," *EURASIP Jour. on Appl. Sig. Proc.*, vol. 2003, no. 11, pp. 1135–1146, 2003.

[13] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on ASSP*, vol. 27, no. 2, pp. 113–120, 1979.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A802-7 (11/18)

大規模マイクロホンアレイを用いた発話方向の実時間推定

Real-time sound source orientation estimation using a 96 channel microphone array

中島 弘史†, 菊池 慶子‡, 醍醐 徹‡, 中臺 一博†, 長谷川 雄二†, 金田 豊‡

Hirofumi NAKAJIMA†, Keiko KIKUCHI‡, Toru DAIGO‡, Kazuhiro NAKADAI†,

Yuji HASEGAWA†, Yutaka KANEDA‡

†(株) ホンダ・リサーチ・インスティチュート・ジャパン

‡東京電機大学工学研究科

†Honda Research Institute Japan Co., Ltd. ‡Tokyo Denki University

nakajima@jp.honda-ri.com

Abstract

This paper addresses sound source orientation estimation using a 96ch microphone array. We proposed a beamforming method with estimation of sound source directivity, and reported orientation estimation of a speech source such as a loudspeaker or an actual human. However, this method, transfer function to design a beamformer required the same as that of the target sound source. Otherwise, the performance deteriorated due to a mismatch between these two transfer functions which was mainly caused by phase errors and outliers. In addition, voice activity detection (VAD) was manually performed. To solve the former, we propose amplitude-based orientation estimation using a histogram. For the latter, we propose two techniques, that is, speech frequency component detection based on inner product, and automatic VAD based on auto-correlation. We constructed a real-time sound source orientation estimation system by introducing our proposed methods. Preliminary experiments showed that sound source orientation estimation with automatic VAD for actual human voices drastically improved even when using a loudspeaker-based transfer function.

1 はじめに

ロボット聴覚の研究分野[1]では，人の聴覚と同様の機能を有する聴覚システムの構築を目指した研究が行われている[2]．主に音源定位（位置推定），音源分離，音声認識といった聴覚機能が研究されている．しかし，人・ロボットインタラクションを考慮すると，これらの機能だけで十分とは言えず，例えば，音源（話者）の向きを推定する機能（以後，発話方向推定と記す）も重要である．図1は，発話方向推定が必要となる例である．図中の女声話者は，ロボットでは無く，別の男性に話しかけている．しかし，発話方向推定を有しないロボットは，女性の発話に反応してしまう．つまり，ロボットはこの女声の発話対象が自分では無い事を判断する必要がある．

画像処理を利用して，顔の向きから発話方向を推定することも可能である．画像処理分野では，顔の検出が大きな研究トピックの1つであり，実時間で高精度に顔の位置と向きを推定可能な手法が報告されている[3]．しかし，画像処理では，周囲が暗い場合や，逆光の場合に推定が難しい．また，処理に必要な計算量が多い事から，ハードウェアが高価になる等の欠点がある．

一方，発話方向を音響処理により，推定しようとする試みは少なく[4]，実時間システムとして構築した例は無い．しかし，人間は目隠しを付けた状態でも，聴覚的な手がかりから発話方向を高精度に知覚できることが加藤らにより報告されている[5]．この報告では，正面にいる話者の発話方向を45°間隔で推定して答えるタスクにおいて，80%以上の正解率が得られている．従って，音響処理により発話方向推定を実現することは，人の聴覚と同様の機能の構築を目指すロボット聴覚の研究として，原理的に重要であると考える．

本研究では，ロボット本体に搭載したマイクロホンからの入力情報に限らず，周囲環境に設置したマイクロホンからの入力情報もロボットが聴覚機能を実現するために利用可能であると考える．我々は，これまでに部屋の壁に設置したマイクロホンアレイを利用した発話方向の推定法を提案した[6]．しかし，このシステムは，(1) 推定精度が低い，(2) 発話区間の検出に手動調整が必要，(3) 実時間性が低いという課題があった．本報告では，これらの課題を解決するため，[A] 振幅特性の抽出処理，[B] 自己相関関数に基づく発話区間検出処理，[C] 周波数マスク処理，[D] ヒストグラム処理を導入した．それぞれの課題と処理の対応は，課題 (1) に対し [A][B][C][D]，課題 (2) に対し [B]，課題 (3) に対し [B][D] となっている．本稿では，これらの処理を導入した実時間音源方向推定システムの構築とその評価結果について報告する．

2 従来の発話方向推定方法とその課題

本章では，筆者らが発話方向の推定法として用いている拡張ビームフォーミングによる音源方向の推定方法について説明する．

Figure 1: An example of importance of sound source orientation estimation

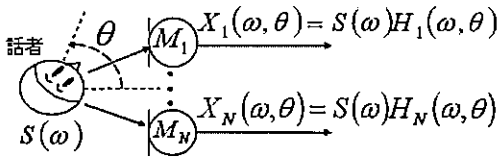


Figure 2: Sound source orientation estimation model

2.1 音源方向への拡張ビームフォーミング

ビームフォーミング（以後，BF と記す）は，空間的な指向性を形成する技術である．BF は，その指向性の焦点を走査することで，音源パワーの空間分布を推定することができ，その最大値から音源位置を推定できる（走査 BF）[7]．BF は，ある特定の位置に対して焦点を形成するよう設計するのが一般的である．筆者らは伝達関数を音源の向きによって変化する関数に拡張することにより，位置だけでなく，音源の向きに対しても焦点を形成する BF が設計できることを示した（拡張 BF）[8]．この拡張 BF を用いると，一般的な走査 BF と同様の処理により，音源の位置だけでなく，音源の向きも推定することができる（拡張走査 BF）．具体的には，それぞれ音源の位置と向きが異なる伝達関数を元に BF を設計し，その出力が最大となる BF の焦点位置と方向を，それぞれ発話者の位置と向きとして推定する．

拡張走査 BF は，BF を遅延和で設計した場合，伝達関数と入力信号の内積の最大値によって方向を推定するのと等価である．従って，拡張 BF の設計，および拡張 BF の走査は，それぞれ位置と向きが異なる複数の伝達関数を集めたデータベースの作成（以後，伝達関数データベースと記す），および伝達関数データベースと入力信号の照合と見なすことができる．本章では簡略化のため，音源の向きのみに対する拡張走査 BF について，データベースの作成と照合の観点から説明する．

2.2 音源の向きに拡張した伝達関数モデル

図 2 に N 素子のマイクロホンアレイを用いた音源方向推定のモデルを示す．$S(\omega)$ は周波数 ω での音源（話者）の周波数特性，M_k は k 番目のマイクロホン ($k=1,2,...,N$)，$H_k(\omega,\theta)$ は話者が θ 方向を向いている時の話者 - マイクロホン間の伝達関数である．ここでは，話者の位置は既知であるとした．マイクロホン M_k での受音信号 $X_k(\omega)$ は，

$$X_k(\omega)=S(\omega)H_k(\omega) \tag{1}$$

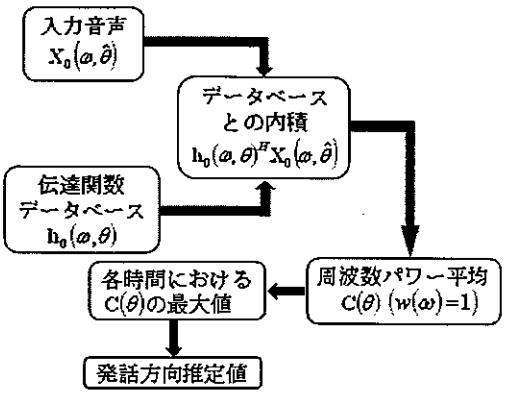


Figure 3: Diagram for conventional method

各変数をベクトルで表現すると

$$\begin{aligned}\mathbf{h}(\omega,\theta) &= [H_1(\omega,\theta),...,H_N(\omega,\theta)]^T \\ \mathbf{X}(\omega,\theta) &= [X_1(\omega,\theta),...,X_N(\omega,\theta)]^T \\ &= [S(\omega)H_1(\omega,\theta),...,S(\omega)H_N(\omega,\theta)]^T \\ &= S(\omega)\mathbf{h}(\omega,\theta)\end{aligned} \tag{2}$$

と表すことができる．

2.3 伝達関数データベース

伝達関数データベースは，各向きの音源から各マイクロホンまでの伝達関数を集めたものである．方向推定においては，伝達関数ベクトル $\mathbf{h}(\omega,\theta)$ のベクトルの向きのみが必要となるため，次式により各周波数および各方向で大きさを 1 に正規化した $\mathbf{h}_0(\omega,\theta)$ を用いた．

$$\mathbf{h}_0(\omega,\theta)=\frac{\mathbf{h}(\omega,\theta)}{\sqrt{\mathbf{h}(\omega,\theta)^H\mathbf{h}(\omega,\theta)}}=\frac{\mathbf{h}(\omega,\theta)}{|\mathbf{h}(\omega,\theta)|} \tag{3}$$

ここで H は複素共役転置を示す．正規化により，伝達関数に含まれる出力機器の特性（スピーカの周波数特性など）を含まない伝達関数を得ることができる．

2.4 データベースを用いた発話方向の推定

図 3 に従来の音源方向推定の処理のブロック図を示す．話者が発話方向 $\hat{\theta}$（未知）に向いて発話した時の受音信号ベクトル $\mathbf{X}(\omega,\hat{\theta})$ は，その大きさを 1 に正規化したものを $\mathbf{X}_0(\omega,\hat{\theta})$ とおけば，

$$\mathbf{X}_0(\omega,\hat{\theta})=\frac{S(\omega)\mathbf{h}(\omega,\hat{\theta})}{|S(\omega)\mathbf{h}(\omega,\hat{\theta})|}=\frac{S(\omega)}{|S(\omega)|}\mathbf{h}_0(\omega,\hat{\theta}) \tag{4}$$

となる．式 (3) と式 (4) の内積の絶対値 $C_\omega(\omega,\theta)$ は

$$\begin{aligned}C_\omega(\omega,\theta) &= \left|\mathbf{h}_0(\omega,\theta)^H\mathbf{X}_0(\omega,\hat{\theta})\right| \\ &= \left|\frac{S(\omega)}{|S(\omega)|}\mathbf{h}_0(\omega,\theta)^H\mathbf{h}_0(\omega,\hat{\theta})\right| \\ &= \left|\mathbf{h}_0(\omega,\theta)^H\mathbf{h}_0(\omega,\hat{\theta})\right|\end{aligned} \tag{5}$$

となり，方向は θ と $\hat{\theta}$ の伝達関数の類似度を示す．この $C_\omega(\omega,\theta)$ を周波数で平均した平均類似度 $C(\theta)$ を

$$C(\theta)=\sum_\omega C_\omega(\omega,\theta) \tag{6}$$

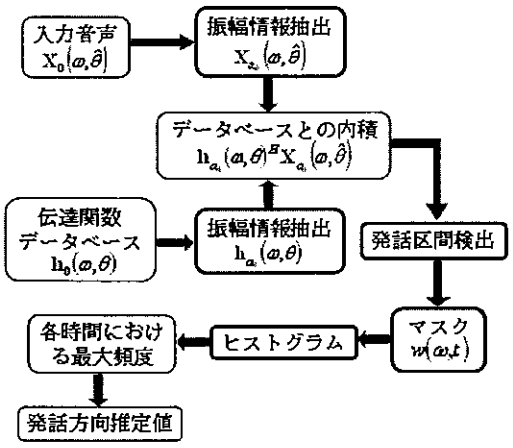


Figure 4: Diagram for proposed method

として計算する．従来法では，この $C(\theta)$ が最大となる θ を音源方向として推定した．

2.5 従来法の課題点

筆者らは実際に，室内壁面に設置した大規模マイクロホンアレイを用いて拡張BFによる人の発話位置・発話方向推定を示した[6]．しかし，このシステムは，(1) 推定精度が低いことが課題であった．この原因について調査した結果，推定精度低下は，主にデータベース作成時と実際の発話時における伝達関数が一致しないことに起因することがわかった．実際にデータベースで用いる伝達関数として推定対象の人間の実発話に基づく伝達関数を用いることで，精度を向上させることができた[9]．しかしこの方法では発話者全員の伝達関数を測定する必要があり実用的ではとはいえない．また，従来法には他にも，(2) 発話区間の検出に手動調整が必要，(3) 実時間性が低いという課題があった．

3 発話方向推定処理の高性能化

従来法の課題点を解決するため，本研究では，スピーカを用いて測定した伝達関数に基づく発話方向推定処理の改良を行うことにより，方向推定処理を向上させる手法を提案する[10]．図4に従来の音源方向推定の処理のブロック図を示す．提案法は，具体的には従来法に対し，新たに [A] 振幅特性の抽出処理，[B] 自己相関関数に基づく発話区間検出処理，[C] 周波数マスク処理，[D] ヒストグラム処理の4つの処理を加えた方法となっている．導入した各処理の課題に対して期待される効果に関しては本章の各節で，実際の有効性に関しては第5章で説明する．

3.1 振幅特性の抽出 [A]

従来の推定方法では，式（6）に示すように，伝達関数の複素成分（振幅成分と位相成分）の内積から発話方向推定を行っていた．しかし，高周波帯域において位相成分は系の変動に敏感で変化しやすい．系の変動は，例えば，話者の口の高さや位置の変化によって起こる．位置の変化量が少ない場合，伝達関数の変動は，特に高周波において位相が大きく変化する．そのため，内積値の抽出の際に，位置の変化に対して比較的ロバストな振幅成分を抽出する処理を加えることで，方向推定精度が向上するものと期待できる．以下，具体的には，振幅成分にもとづく内積計算の方法について定式化する．伝達関数の振幅成分をベクトル化したものを

$$\mathbf{h}_a(\omega,\theta) = [|H_1(\omega,\theta)|, ..., |H_N(\omega,\theta)|]^T \tag{7}$$

とし，これを正規化したものを

$$\mathbf{h}_{a0}(\omega,\theta) = \frac{\mathbf{h}_a(\omega,\theta)}{|\mathbf{h}_a(\omega,\theta)|} \tag{8}$$

とする．同様に，各受音信号の振幅成分をベクトル化したものを

$$\mathbf{X}_a(\omega,\theta) = [|X_1(\omega,\hat{\theta})|, ..., |X_N(\omega,\hat{\theta})|]^T \tag{9}$$

とおき，これを正規化したものを

$$\mathbf{X}_{a0}(\omega,\theta) = \frac{\mathbf{X}_a(\omega,\hat{\theta})}{|\mathbf{X}_a(\omega,\hat{\theta})|} \tag{10}$$

とする．式（6）と同様に内積値の絶対値 $C_{a\omega}(\omega,\theta)$ は

$$C_{a\omega}(\omega,\theta) = \left|\mathbf{h}_{a0}(\omega,\theta)^H \mathbf{X}_{a0}(\omega,\hat{\theta})\right|$$

として計算され，これをもとにに平均類似度 $C_a(\theta)$ を

$$C_a(\theta) = \sum_{\omega} w(\omega,t) C_{a\omega}(\omega,\theta) \tag{11}$$

として計算し，この $C_a(\theta)$ が最大となる方向 θ を発話方向の推定値とする．ここで $w(\omega,t)$ は時間-周波数マスクであり，提案法で新たに導入したものである．本稿では，$w(\omega,t) = w_\omega(\omega)w_t(t)$ として時間と周波数のマスクを独立に計算した．それぞれの計算法については，$w_t(t)$ は [B] 発話区間検出処理，$w_\omega(\omega)$ は [C] 周波数マスクの節で述べる．方向推定処理は，時間フレーム単位で行われるため，入力信号 $\mathbf{X}_0(\omega,\hat{\theta})$，内積の絶対値 $C_\omega(\omega,\theta)$ は時間の関数でもあるが，簡略化のため変数 t は省略した．

3.2 発話区間検出 [B]

従来法では，入力信号のレベルに対し，手動調整で定めた閾値を用いて発話区間検出を行っていた．しかし，入力信号レベルは，話者や発話方向の向きによって異なるため，複数の話者や発話方向の向きに対して，高精度に発話区間を検出するためには，各条件で閾値を手動で調整する必要があり，現実的な処理とはいえなかった．提案法では，入力信号のレベルではなく，入力信号の自己相関関数を利用した発話区間検出方法を導入した．この方法は音声の母音の周期性を根拠とするもので，音声のレベル変化によらず発話区間を検出できるという利点があるだけでなく，非周期性雑音に対しても頑健であり，S N比の低い環境で有効に動作することが期待できる．具体的には，まず，基準とするチャンネルで受音した信号を分析窓長（1024点）で切り出し，その自己相関関数 $\phi(\tau)$ を計算する．信号が周期性を有する場合，$\phi(\tau)$ はその周期に対応する τ にピークを有する．$\phi(\tau)$ が最大となる $\tau = 0$ のピークを除き，$\tau > 0$ の範囲における $\phi(\tau)$ のピークの最大値が閾値（α と設定）を超えた場合に，周期性を持つ信号であるとし，発話区間と判定した．なお，本稿では，雑音による小さいピークやディップを除外するため，$\phi(\tau)$ が β 以下

の値をとった時間以降の最大値を「$\tau > 0$ の範囲における $\phi(\tau)$ のピークの最大値」と判定した．この処理で判定された音声区間から 2 値の時間マスク $w_t(t)$ を生成した．

$$w_t(t) = \begin{cases} 1 & \text{発話区間} \\ 0 & \text{非発話区間} \end{cases} \quad (12)$$

パラメータ α および β は実験的に定め，本稿では $\alpha = 0.5$,$\beta = -0.2$ とし，すべての条件や発話でこの値を利用した．

上記の処理を用いて予備的な実験を行った結果，音声の残響部分では，正しい発話方向が得られないことが明らかになった．これは残響音が音源方向とは異なる方向から到来するためであると推定される．このため提案法では，上記の処理に加えて音声の残響部分を発話区間から除外する処理を導入した．最適な除外方法の検討は今後の課題とし，今回は検出した発話区間の最後尾 2 フレーム（時間データ 1,024 点に相当）を除外したものを最終的な発話区間とした．

3.3 周波数マスク [C]

従来法では，音声の周波数特性を考慮せず，全ての周波数帯域で一律に平均した平均類似度の最大値から発話方向を推定していた．この方法では，音声がほとんど含まれず，正しい方向が得られない周波数帯域の成分も含まれるため，推定精度が低下する．提案法では，これを防ぐために類似度の方向に対する最大値 $C_{max}(\omega)$ に基づく 2 値の周波数マスク $w_\omega(\omega)$ を生成した．

$$w_\omega(\omega) = \begin{cases} 1 & if\ C_{max}(\omega) > \overline{C_{max}(\omega)}\delta \\ 0 & otherwise \end{cases} \quad (13)$$

ここで $\overline{A}$ は A の時間平均，δ は閾値である．今回は，δ は実験的に 1.5 と定めた．

3.4 ヒストグラム [D]

従来法では，各周波数 ω，および各方向 θ で算出した方向類似度を周波数方向に平均した $C(\theta)$ から発話方向を推定している．しかし，方向類似度（内積値）の絶対的大きさは周波数に依存するため，平均の結果も絶対的な内積の大きさで加重される．これに対して，周波数マスクで音声成分と判定された周波数成分については，均等加重で評価を行うという方法が考えられる．具体的には，平均した方向類似度 $C(\theta)$ ではなく，各周波数について方向類似度が最大となる方向を算出し，その数をカウントしたヒストグラム $C_{Hist}(\theta)$

$$C_{Hist}(\theta) = \sum_{\omega} w(\omega,t)U(\omega,\theta) \quad (14)$$

$$U(\omega,\theta) = \begin{cases} 1 & if\ \theta = \mathrm{argmax}_\theta[C_{a\omega}(\omega,\theta)] \\ 0 & otherwise \end{cases} \quad (15)$$

$$(16)$$

が最大値を取る θ を発話方向の推定値とした．ここで $\mathrm{argmax}_\theta[]$ は，括弧内の関数が最大となる θ を示す．

4 発話方向の実時間推定システム

提案法をもとに実時間で動作する発話位置と発話方向の推定システムを開発した．図 5 に開発したシステムのブロック図を示す．各ブロックの詳細は下記の通りである．

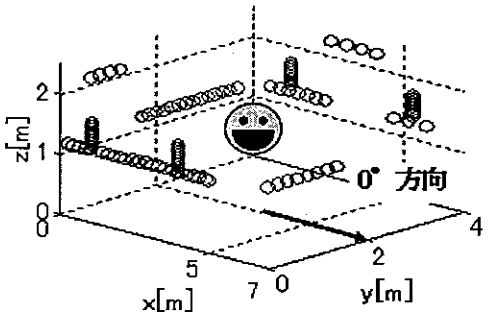


Figure 5: Real-time location and orientation estimation system

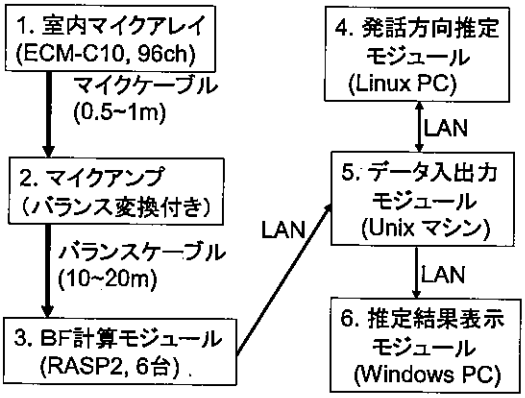


Figure 6: Arrangement of microphones

4.1 室内マイクロホンアレイ (1)

室内マイクロホンアレイは，実験室の壁面に分散して埋め込まれた 96 個のマイクロホンで構成した．実験室は広さ 7m × 4m，高さ 3.5m，残響時間が約 230ms で，壁三面が吸音素材，一面がガラス面である．図 6 は，マイクロホンの配置を示しており，○印の位置にマイクロホンが配置されている．マイクロホン配置は，本研究の目的に特化させたものではなく，水平および垂直方向に概ね等間隔に配置したアレイを周囲四壁に分散した配置となっている．マイクロホンは，SONY ECM-C10 である．このマイクロホンは，無指向性で，民生品の中では比較的 SN 比や安定度が高い．各マイクロホンは，0.5〜1m 程度の短いケーブルでマイクアンプに接続することにより，環境電磁波によるノイズを防ぐ工夫を行った．

4.2 マイクアンプ (2)

マイクアンプは，日東紡音響エンジニアリング社の AEMM-04 を 24 台利用している．このアンプは，1 台で 4ch のマイク電源の供給，60dB〜80dB のアンプ，インピーダンスおよびバランス変換を行うことができる．出力は，シールドされたバランスケーブルで室外の入力換器まで接続されている．ケーブルの距離は 10〜20 mあるが，信号レベルが大きく，バランスで伝送されているため，このケーブルによる信号の劣化はほとんどない．

4.3 BF 計算モジュール (3)

BF 計算モジュールは，日本電子システムテクノロジー社の RASP2 を 6 台利用した．RASP2 は，16ch の AD 変換

と汎用 CPU が内蔵されており，OS として Linux が動作する．この RASP2 を利用して，走査 BF による音源定位と振幅スペクトルにもとづく方向推定用の拡張走査 BF の出力計算 [A] を行っている．

音源定位は，室内の探査空間 (4m × 3m) を水平面内に 25cm 間隔で離散化した計 221 点に焦点をもつ BF を設計し，その BF の出力が最大となる点の位置を音源位置として推定した．音源定位 BF は，計算量削減のため，入力信号スペクトルの中で SN 比の高い帯域を 5 帯域選択し，その帯域についてのみ計算している．

方向推定は，定位で得られた位置についてのみ 45° 刻みで離散化した 8 方向の音源方向についての拡張 BF を計算し，その各方向の BF 出力値と定位結果をモジュールに送信する．計算量の制限のため，拡張 BF で計算する周波数帯域は，0〜1kHz である．

4.4 データ入出力モジュール (4)

データ入出力モジュールは，我々が独自に開発したミドルウェアである MMI Ver.2[11]を利用した．このモジュールを起動することで，さまざまな種類のデータを非同期にネットワークを介してアクセスできる．本システムでは，方向推定用 BF の結果と，VAD，ヒストグラムを行った方向推定結果のデータアクセスに利用している．

4.5 発話方向推定モジュール (5)

本モジュールは，データ入出力モジュール (MMI) から，定位および各方向の BF 出力値を取得し，最終的な推定位置と方向を計算して，MMI に送るプログラムである．本プログラムでは，各フレーム時刻で送信される BF 出力値から，時間一周波数マスク [B][C] の生成と適用，ならびに方向類似度のヒストグラム [D] 計算を行っている．ヒストグラムは，発話区間と判定された時間フレームが 4 フレーム蓄積された時点で，ヒストグラムが最大となる方向を定位方向として MMI に送信する．また本モジュールでは，より信頼性の高いデータを抽出するために，時間一周波数マスクに加えて，定位結果の不連続性が大きい場合にもデータを除外する処理も付加的に行っている．

4.6 推定結果モジュール (6)

本モジュールは，データ入出力サーバから送信される，推定位置および方向の結果をグラフィカルに表示するものである．本モジュールでは，OpenSceneGraph[12]を利用して実装した MMI 用の表示プログラム MMI Viewer を利用した．図 7(a) は本システムの利用状況，図 7(b) は Viewer の表示例である．本システムを利用して，実際に発話者の位置と方向を実時間で推定可能であることが確認できた．

5 評価実験

提案法の有効性を示すための評価実験を行った．実験では，提案法と従来法による推定精度の比較のため，実験条件は両手法で処理可能な条件に設定した．そのため今回は，位置推定モジュールの誤差や周波数帯域制限に関する方向推定誤差の評価は行わなかった．本章では，はじめに実験で用いた伝達関数データベース，評価用音声について記述し，各処理過程の評価，推定精度の評価について述べる．

(a) 本システム利用状況

(b) Viwer の表示例

Figure 7: Speaker and viewer

5.1 伝達関数データベースの作成

伝達関数データベースは，スピーカを音源として測定したインパルス応答を FFT することにより作成した．スピーカは，GENELEC 社の 1029A を用いた．スピーカは部屋の中央 (x=3m, y=2m) に配置し，図 6 の 0° 方向から反時計回りに 15° 刻みで 345° まで回転させた（計 24 方向）．インパルス応答収録時のサンプリング周波数は 16kHz，音源信号は，信号長 2^{14} の TSP 信号とした．伝達関数は，このインパルス応答の初期部分（1,024 点）を切り出し，FFT を行うことにより計算した．

5.2 評価用音声の収録

伝達関数データベースの作成時に利用した実験室で行った．話者は男性 1 名で，部屋の中央 (x=3m, y=2m) で 0°, 90°, 180°, 270° の 4 方向に向き，「あ，い，う，え，お」と発話した音声を収録した．

5.3 各処理過程の評価

提案法の各処理が適切に動作していることを確認するため，はじめに発話区間検出処理についての評価を行った．図 8 に，受音した信号のスペクトログラム，図 9 に発話区間検出処理結果を示す．図 8 より音声成分は 8kHz まで含まれており，500Hz 以下の低周波域には雑音が存在すること，音声区間の末尾は残響によるレベルの低い部分が含まれることがわかる．図 9 を見ると，このような雑

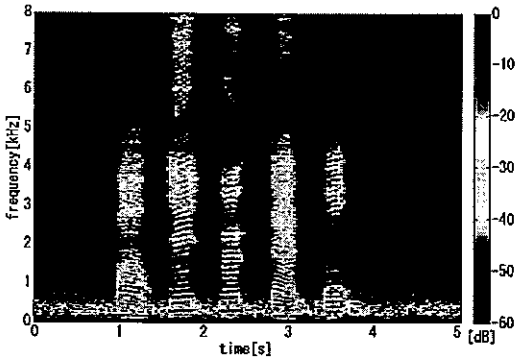


Figure 8: Input signal spectrogram

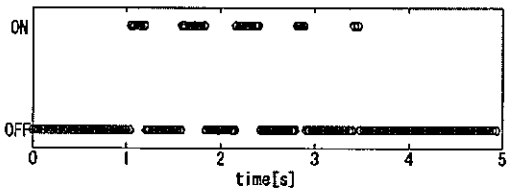


Figure 9: Voice activity detection result

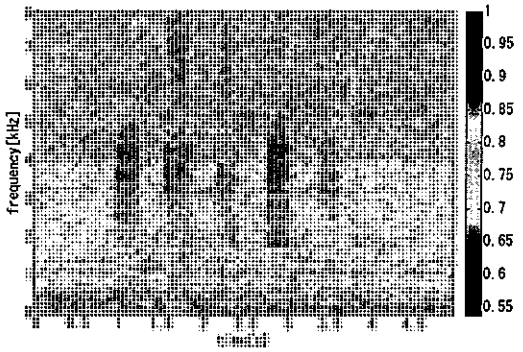


Figure 10: Maximal value of the orientation similarity in time-frequency domain

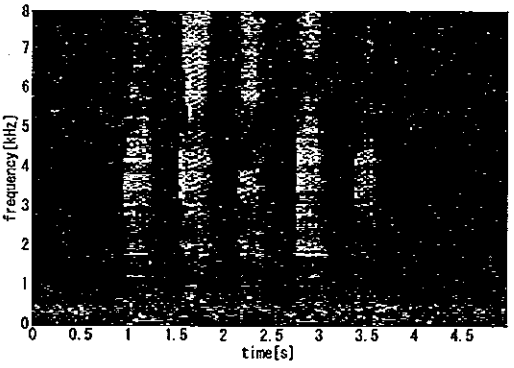


Figure 11: Time-frequency mask

音・残響を含む音声信号に対しても，発話区間が検出できることがわかる．また図 9 で検出された発話区間が，図 8 のレベルの高い区間よりも短い原因は，残響が顕著な末尾の部分を除く処理を行っているためである．

つぎに，周波数マスク処理についての評価を行った．周波数マスクは，3.3 節で述べたとおり，方向類似度の最大値 $C_{max}(\omega)$ を元に生成している．図 10 は，発話方向推定の元となる各時間および各周波数における方向類似度の最大値を表した図である．図 10 を図 8 と比較すると，2kHz 以上の高周波数域では，音声の存在する区間で内積値は大きくなっていることがわかる．しかし 1kHz 以下の低周波数帯域では，音声区間と非音声区間によらず内積の値に差が見られず，この帯域を利用することは誤差の原因となる可能性が考えられる．

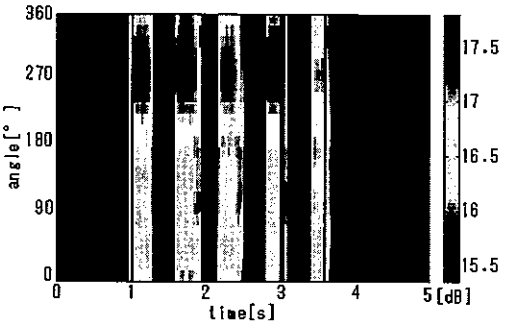


Figure 12: Orientation similarity for each angle

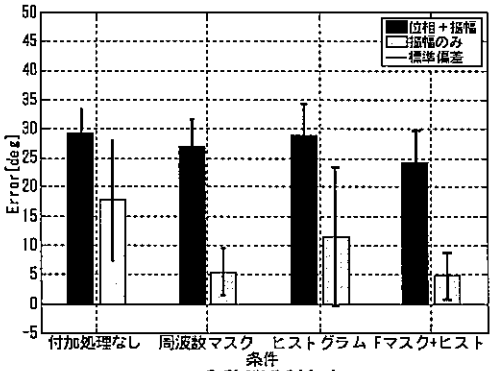
図 11 は，音声区間検出処理と周波数マスク処理から得られる最終的な時間－周波数マスクであり，黒い箇所がマスクされる部分である．マスクが，単に音声が強い部分を抽出しているのではなく，方向類似度の差が高く，方向推定に有効な高周波成分を多く抽出できることがわかる．

図 12 は，時間周波数マスクを施し，最終的に方向を算出する前の方向類似度－時間特性である．実際の音源方向は 270° であり，概ね正しい方向で方向類似度が高い値をとっていることがわかる．

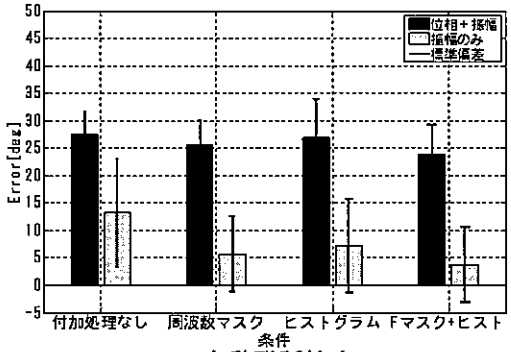
5.4 推定精度の評価

提案法の有効性を示すため，従来法と提案法の発話方向推定結果の誤差を評価した．また提案法で導入した 4 つの処理（[A] 振幅特性の抽出，[B] 発話区間検出，[C] 周波数マスク，[D] ヒストグラム）の寄与度について検討するため，それぞれの処理を導入した場合と導入しない場合（計 1 6 通り）の処理について評価した．全ての処理の導入しない処理が従来法であり，全ての処理を導入した処理が提案法である．

図 13 は，各処理で推定した発話方向推定結果の誤差と標準偏差を示している．図 13（a）は，[B] の発話区間検出を導入せず（手動調整）により得られた発話区間を用いた結果，図 13（b）は提案法で導入した自動発話区間検出を用いた結果である．各図において濃色の棒は，[A] の振幅特性の抽出を導入しなかった場合（振幅と位相を用いた内積処理）の結果，淡色の棒は [A] 振幅特性の抽出を導入した場合の結果を表している．各図の棒グラフは，左から周波数マスクおよびヒストグラム処理なし，周波数マスク [C] のみ導入，ヒストグラム [D] のみ導入，周波数マスクとヒストグラム処理 [C][D] を導入した場合となっている．図 13(a) と (b) を比較して，これら 2 つはほぼ同程度の誤差の大きさとなっている．このことより，提案する自動発話区間検出法は，手動調整とほぼ同程度の性能を達成できることが確認された．次に，各図における濃色棒と淡色棒を比較すると，すべての条件で淡色棒の方が誤差が小さくなっていることがわかる．これより，「振幅と



(a) 手動発話検出



(b) 自動発話検出

Figure 13: Estimation error and standard deviation

位相の両方」を用いる処理（濃色棒）より「振幅のみ」を用いた処理（淡色棒）が優位であることが示された．次に図 13（a）（b）における 4 条件を比較してみると，いずれの図でも「周波数マスクとヒストグラムの両方の処理を行った場合」が最も誤差が小さい．このことより，今回改良のために提案した周波数マスクとヒストグラムの 2 つの処理は，共に推定誤差を低減する効果があることが確認できた．従来法である，図 13（a）の「付加処理なし」の「位相＋振幅」の条件では，平均誤差は約 30° であるのに対して，改良法である図 13（b）の「周波数マスクとヒストグラムの両方の処理」を用いると，平均誤差が約 4°，標準偏差が約 7° と大幅な性能改善を達成することができた．また，この誤差の絶対的な大きさも，今回のデータベースが角度 15° おきのものであり，推定結果も 15° 間隔で求めたことから，推定の量子化角度以内の誤差であったと評価できる．

6 まとめ

本報告では，発話方向を高精度に推定できる手法を提案した．提案法は，拡張 BF に基づく従来の方向推定法に対し処理の追加と改良を行うことにより，高精度な方向推定を実現した．また提案法を元に実時間で動作する発話方向推定システムを開発した．推定誤差を評価した結果，従来法で 20° 以上あった方向推定誤差が，提案法では 5° 以下に低減できることがわかった．マイクロホン数の削減や主要周波数の選択などによる計算量の削減，画像情報の取得によるロバスト化などが今後の課題である．

参考文献

[1] K. Nakadai, T. Lourens, H. Okuno, and H. Kitano, "Active audition for humanoid," in *17th National Conf. on Artificial Intelligence (AAAI2000)*. AAAI, 2000, pp. 832–839.

[2] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "High performance sound source separation adaptable to environmental changes for robot audition," in *2008 International Conference on Intelligent Robots and Systems (IROS2008)*. IEEE/RSJ, 2008, pp. 2165–2171.

[3] G. Dedeoglu, T. Kanade, and S. Baker, "The Asymmetry of Image Registration and its Application to Face Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 807–823, 2007.

[4] P. C. Meuse and H. F. Silverman, "Characterization of talker radiation pattern using a microphone array," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1994, pp. 257–260.

[5] 加藤 宏明, 竹本 浩典, 西村 竜一, "聴覚手がかりによる話し手の向きの知覚," *2008* 年 *8* 月 *5* 日 応用音響研究会 当日配布資料, 電子情報通信学会, 2008.

[6] K. Nakadai, H. Nakajima, K. Yamada, Y. Hasegawa, T. Nakamura and H. Tsujino, "Sound Source Tracking with Directivity Pattern Estimation Using a 64ch Microphone Array," in *2005 International Conference on Intelligent Robots and Systems (IROS2005)*. IEEE/RSJ, 2005, pp. 192–202.

[7] 菊間 信良, "アレーアンテナによる適応信号処理," 科学技術出版, 1999.

[8] 中島 弘史, "音源の方向を推定可能な拡張ビームフォーミング," 日本音響学会秋期研究発表会, 日本音響学会, 2005, pp. 619–620.

[9] 醍醐 徹, 菊池 慶子, 中島 弘史, 中臺 一博, 長谷川 雄二, 金田 豊, "室内残響を考慮した大規模マイクロホンアレイによる発話方向の推定," 日本音響学会秋期研究発表会, 日本音響学会, 2007, pp. 627–630.

[10] 菊池 慶子, 醍醐 徹, 中島 弘史, 中臺 一博, 長谷川 雄二, 金田 豊, "大規模マイクロホンアレイによる発話方向推定の検討," 信学技報 *EA2008-37*, 電子情報通信学会, 2007, pp. 13–18.

[11] 鳥井豊隆, 長谷川雄二, 中野幹生, 中臺一博, 辻野広司, "人・ロボットインタラクションシステムの為のミドルウェアの開発," 第 *7* 回システムインテグレーション部門講演会 *(SI2006)*, 計測自動制御学会, 2006, pp. 610–611.

[12] http://www.openscenegraph.org/projects/osg

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A802-8 (11/18)

パーティクルフィルタリングによる移動ロボットからの二次元音源地図作成

加賀美 聡[1,2,3] Simon Thompson[1,3] 佐々木 洋子[2,1]
溝口 博[2,1] 榎本 格士[4]

1 産業技術総合研究所デジタルヒューマン研究センター
2 東京理科大学理工学部機械工学科
3 JST, CREST
4 関西電力（株）

Abstract

本論文は移動ロボットに搭載したマイクアレイの短時間の方位角計測の結果から、パーティクルフィルタを用いて2次元の音源地図作成を行う手法について述べる。本手法では音源は方位角と距離のそれぞれに独立したガウス分布に従うと仮定したモデルによりパーティクルフィルタの分散を決定している。

これまで我々が作成してきたビームフォーミング用の32ch低サイドローブマイクアレイと、帯域選択法 (FBS) を組み合わせた音源定位手法を用いて、移動ロボットにより実験を行い、提案手法の有効性を確認した。

1 Introduction

A sound source mapping function is vital for a robot that operates in a human environment. Bearing only Simultaneous Localization and Mapping (SLAM) technique is actively investigated in last several years, basically by using optical cameras (ex. [1]). However sound signal has significant difference in two points, directional localization and characteristics of the source. Difficulty for directional localization is coming from acoustic reverberation, diffraction, resonance, interference, and so on. On the other hand, difficulty for the characteristics is that sound generated by source is usually unknown and always changing in time or even sometimes missing.

Particle filters are widely used in perception area in robotics to handle noisy input to estimate surrounding map and/or robot location. Several methods have been proposed by using particle filtering to achieve directional localization and in tracking in microphone centric coordinates(ex. [2, 3]). As for a mapping function, Nakadai et al.[4] presents a method to map sound source location using a particle filter from a microphone array attached both in a room and on a robot. In this paper, we propose a method to achieve 2D mapping by using only onbody microphone array.

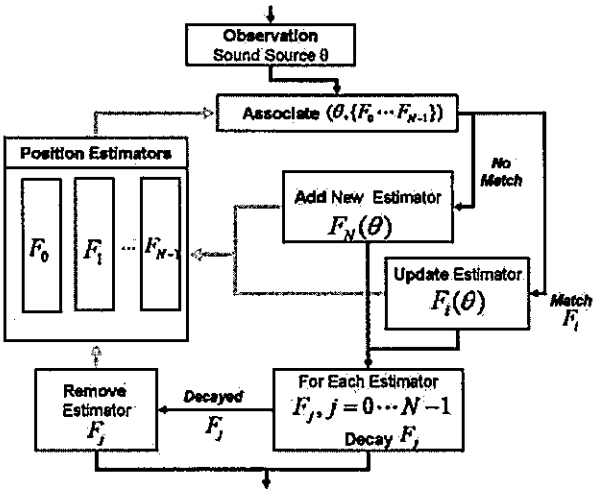


Figure 1: Flow chart of estimator management for 2D sound source mapping with mulitple sound sources.

2 2D Sound Source Mapping

Two dimensional sound source position estimation is achieved by applying a bearing only state estimation technique. Individual particle filters are used to maintain position estimates of a particle sound source, with the set of particles representing a distribution over the x,y coordinate frame. An unknown number of sound sources can be present in a given environmnet, so the number of estimators must be managed, with sound source estimators being created and deleted as required. Figure 1 shows the flow chart of estimator management.

In addition to this, sound sources also emit signals intermittently, so the activity of sound sources over time needs to be monitored. The estimator monitors activity of a particular sound source by use of a decay mechanism. Signal detection causes 'growth' of the decay value, while an absence of signal causes decay. Once a sound source decays to a given value, the estimator is deleted.

Upon observation of new sound source Obs at time k in a particular direction θ_k, a new particle filter estimator F_N is created and initialised from the current robot location with its particles spread over a 2D Gaussian distribution over the direction estimate θ_k, at a default distance r_D. That variance associated with the distributions and are determined by the error in the directional sound source estimate σ_θ, and a default large variance in distance σ_r, reflecting the absence of distance information in the bearing only observation.

Initiatization, then occurs as follows

1. Sound source $Obs(k) = \theta_k$
2. From robot pose $(x, y, \theta)_R$ initialize particles
   $S = \{s_0, \ldots, s_{N_P-1}\}$
   For all s_i in S do
   - $r_i = r_D + G(\sigma_r)$, $\alpha = \theta_k + G(\sigma_\theta)$,
     where $G(\sigma)$ is a function returning a Gaussian distributed random value with variance σ^2
   - $s_i = ((x_R + r_i)\cos(\alpha + \theta_R),$
     $(y_R + r_i)\cos(\alpha + \theta_R))$

The filter then propagates particles representing the probability density function of the sound source location as follows:

1. Observe sound source $Obs(k) = \theta_k$ from $(x, y, \theta)_R$
2. Disperse S, $s_i(k) = s_i(k-1) + \omega$, where ω is a random motion
3. Measure S, such that $p(s_i(k)) = SM(s_i(k), \theta_k\,(x, y, \theta)_R)$
4. Resample S with replacement, based on $p(s_i(k))$

where $SM(s_i(k), \theta_k\,(x, y, \theta)_R)$ is a sensor model returning the probability of making observing a sound source at position $s_i(k)$ at angle θ_k from the current robot position.

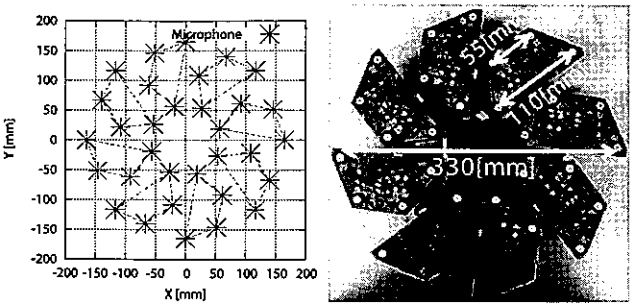


Figure 2: 32ch microphone arrangement (left) and photo (right)

3 Directional Localization

Proposing sound source mapping method can handle noise of directional localization. However, noise should be stochastically small. Therefore, directional localization system needs to be robust from false positive detection. For this purpose, we have been designing and developing low side-lobe microphone arrays that is optimized for Delay and Sum Beam Forming (DSBF) method.

3.1 32ch Low Side-lobe Microphone Array

In order to detect sound source direction for audio input with an unknown frequency, we developed a microphone array and firewire interface board.

The diameter of the microphone array is limited to 33cm due to our mobile robot size. Through simulation of sound pressure distribution, we empirically decided the microphone arrangement to minimize side-lobes. Fig.2(left) shows the resulting microphone arrangement which consists of the octagonal arrangement of eight 4ch microphone boards that have an isosceles trapezoid shape. Fig.2(right) shows the picture of it. The system has 16bits in simultaneous 16khz sampling rate.

Fig.3 shows the beam forming simulation results at 1000, 1400, 2000 [hz]. At each frequency, the focus direction gain compared to side-lobe is 12[dB] at minimum and 16[dB] in average (from 700-2500[hz]).

Fig.4 shows simulated and measured directivity pattern of this microphone array. Horizontal axis is direction, and array is focusing on 0[deg] direction. Vertical axis is signal gain in [dB] compared to focused direction.

3.2 Frequency Band Selection Method

DSBF method has limited performance, especially the method does not remove other signals perfectly (just reduces). Thus, we apply the FBS method[5] after DSBF for the detection of multiple sound sources. FBS is a kind

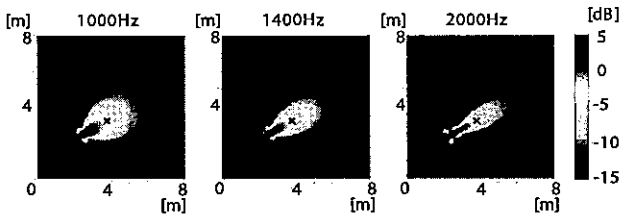


Figure 3: Beamforming simulation result at 1000,1400,2000hz

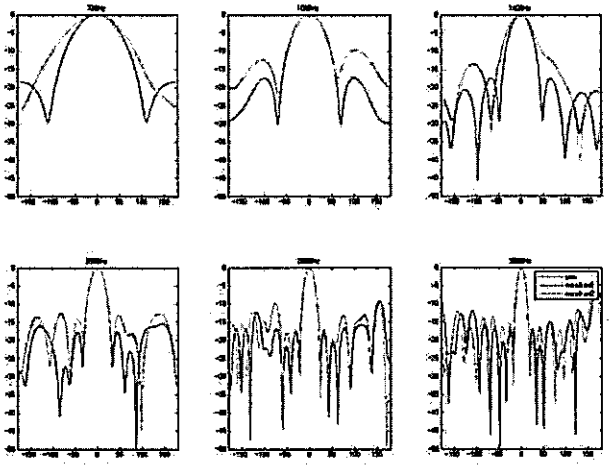


Figure 4: Simulated (red) and measured (green&blue) directivity pattern at our office

of binary mask and segregates objective sound sources from mixed sound by selecting the frequency components judged to be from a common objective sound source.

The process is as follows. Let $X_a(\omega_j)$ and $X_b(\omega_j)$ be the frequency components of DSBF-enhanced signals for objective and noise sources, respectively. The selected frequency component $X_{as}(\omega_j)$ is expressed as Equation(1):

$$X_{as}(\omega_j) = \begin{cases} X_a(\omega_j) & \text{if } X_a(\omega_j) \geq X_b(\omega_j) \\ 0 & \text{else} \end{cases} \quad (1)$$

This process rejects the attenuated noise signal from the DSBF-enhanced signal. The segregated waveform is obtained by the inverse Fourier transform of $X_{as}(\omega)$.

When the frequency components of each signal are independent, FBS can separate the desired sound source. This assumption is usually effective for human's voice or every day sound within a short time period.

Fig.5 shows a procedure. The first step filters out the average signal of each microphone (no delayed signal) input by FBS and finds the loudest sound from the spatial spectrum. When the frequency component of the average signal is higher than any DSBF-enhanced signal

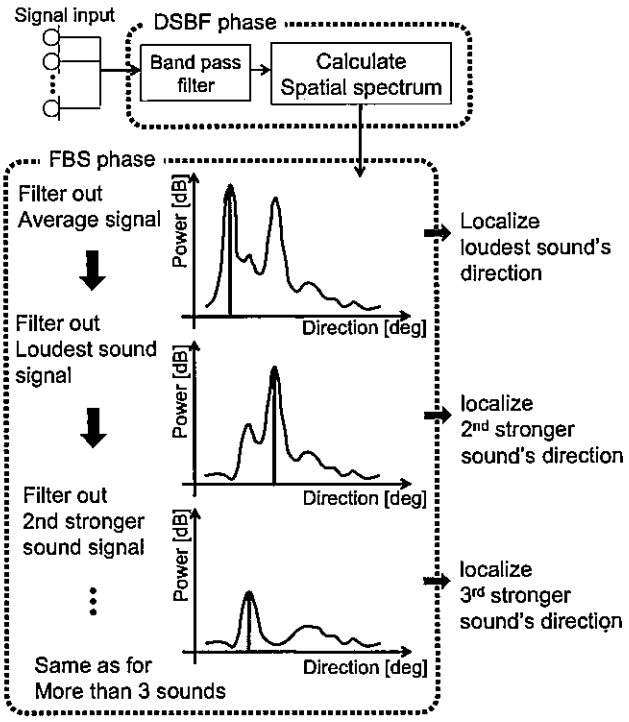


Figure 5: FBS sound localization process

from each direction, the system filters out the spectrum of that frequency. This process rejects omni-directional noise sounds.

The second step filters out the 1st sound signal by FBS, and finds the second strongest sound from the spectrum. When the frequency component of the DSBF-enhanced signal of the 1st sound's direction is higher than that of any other direction, the system filters out the spectrum at each frequency.

If there are more than two sounds, the system finds the third strongest sound, and so on, after filtering out the second strongest sound signal. The method localizes multiple sounds from the highest power intensity to the lowest at each time step. Then the system can continuously localize multiple sound sources and separate each sound source during movement.

4 Experiments

We conducted two experiments using our mobile robot "Pen2"(Fig.6). In experiment 1, four speakers at the robot microphone level setting are used. In experiment 2, five speakers in different height setting are used. Arrangement in experiment 2 is shown in Table 1 and Fig.7.

A commercial motion capture system (Motion Analysis Eagle) with 12 cameras measures robot position in 240[hz] as a ground truth. Standard deviation of robot position measured by this MOCAP system is 0.042[mm]

Table 1: Speaker arrangenet in experiment 2

x[mm]	y[mm]	z[mm]
3,000	500	570
-1,000	-3,000	850
-3,500	-1,500	1,200
670	-1450	2,040
240	1070	1,640

in translation and 1.09e-5[deg] in rotation.

The microphone array locates sound directions at around 12[hz]. Reverberation time T_{60} was 500[msec], and back ground noise level was 50[dBA] (mainly fan noise). Signal noise ratio was 20[dBA] for experiment 1, and 15[dBA] for experiment 2. Sound sources were music, male and female voices.

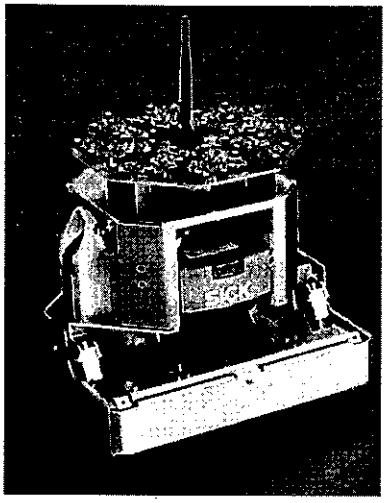


Figure 6: Mobile Robot "Penguin2"

Fig.8 shows the results of localising four sound sources. In this experiment, all the loud speakers are placed on microphone array level. Fig.8(b) shows the convergence of the localisation process and the remaining error. After 100 samples (about 8[s]), the system achieves 2D mapping with around 50[cm] remaining error.

Fig.9 shows an experiment with five sound sources. In this case, sources are placed in different height (from 57 to 204[cm]). One source at (67, -145, 204)[cm] is not found at all. It may be placed too high up and bacause system only conducted directional localization around. The remaining four sources are found. Fig.9(b) shows basically the same kind of convergence performance as the previous experiment. Interestingly, some sources are lost and refound as the robot moves throughout the environment. At each time a sound source is found, convergence occurs as like before.

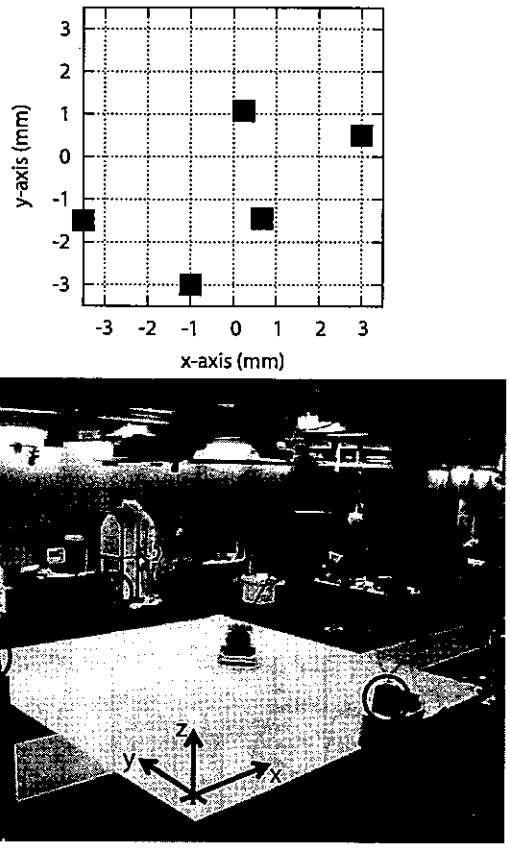


Figure 7: Speaker setting and pictures for experiment 2

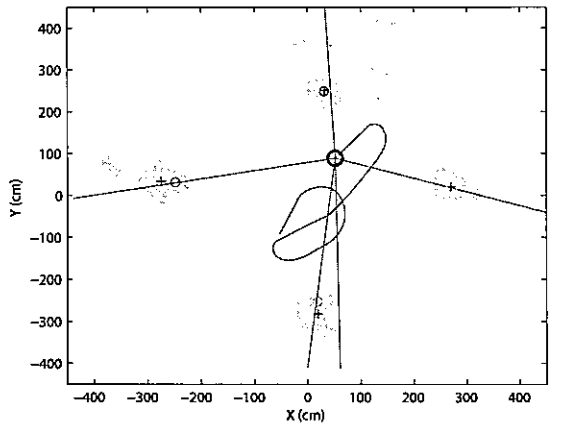
5 Conclusion

This paper proposed a 2D sound source mapping method while robot is in motion, by applying particle filter technique. The method is general for any directional localization. Combined with our 32ch low side-lobe microphone array and with Delay and Sum Beam Forming (DSBF) + Frequency Band Selection (FBS) methods, the system can map 2D arrangement of sound sources. Experimental results show after 100 sampling, detected sound source locations converge in less than 50[cm].

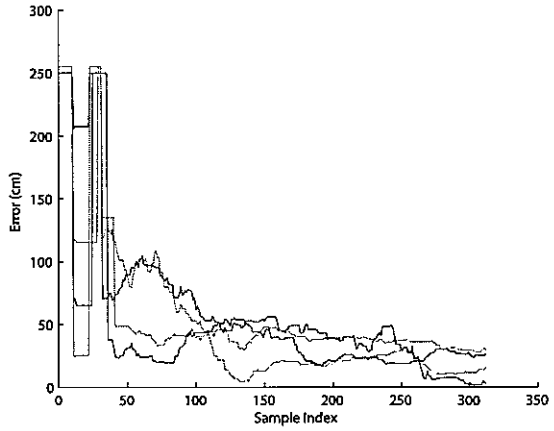
Since one sound source that is located high above the array height is not mapped well in experiment 2, in the future, we would like to 1) extend our mapping function into 3D, 2) optimize microphone array design for two directional localization, 3) more robust and two directional sound source detection method.

参考文献

[1] Sebastian Thrun, Wolfram Burgard, and Dieter Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*, The MIT Press, September 2005.

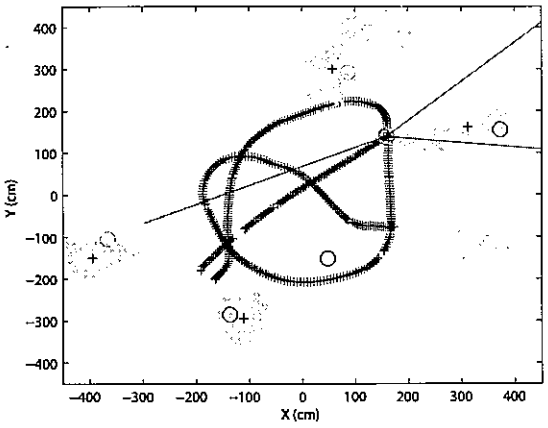


a)

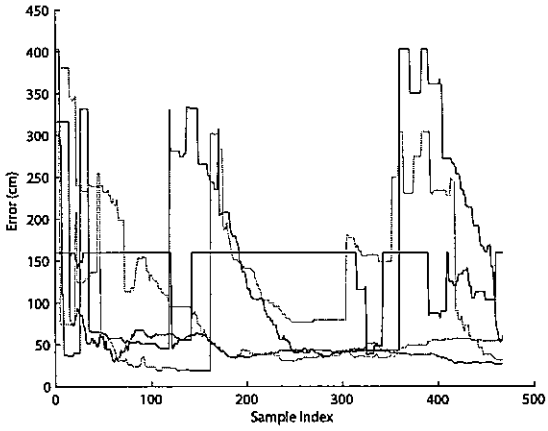


b)

Figure 8: Experiment 1. a) 2D mapping of four sound sources. True position (∘), estimated position (+) and associated particles (yellow dot) that are converged around sound sound sources, are shown. Blue trajectory shows robot motion. b) Sound source mapping error. Distance from true speaker location to each closest estimated sound source over time.



a)



b)

Figure 9: Experiment 2. a) 2D mapping of five sound sources. b) Sound source mapping error

[2] H. Asoh, I. Hara, and H. Asano, "Tracking human speech events using a particle filter," in *In Proceedings International Conference on Audio, Speech and Signal Processing*, 2005, pp. II/1153–1156.

[3] Jean marc Valin, Francois Michaud, and Jean Rouat, "Robust 3d localization and tracking of sound sources using beamforming and particle filtering," in *In Proceedings International Conference on Audio, Speech and Signal Processing*, 2006, pp. IV/224–227.

[4] Kazuhiro Nakadai, Hirofumi Nakajima, Masamitsu Murase, Satoshi Kaijiri, Kentaro Yamada, Takahiro Nakamura, Yuji Hasegawa, Hiroshi G. Okuno, and Hiroshi Tsujino, "Robust tracking of multiple sound sources by spatial integration of room and robot microphone arrays," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2006*, Toulouse, France, May 2006, pp. IV/929–932.

[5] Mariko Aoki, Manabu Okamoto, Shigeaki Aoki, Hiroyuki Matsui, Tetsuma Sakurai, and Yutaka Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoustical Science and Technology*, vol. 22, no. 2, pp. 149–157, 2001.

〒162 東京都新宿区津久戸町 4-7 OS ビル 402 号室 03-5261-3401 Fax: 03-5261-3402

(本研究会についてのお問い合わせは下記にお願いします.)

SIG-AI-Challenges home page (WWW):
http://winnie.kuis.kyoto-u.ac.jp/SIG-Challenge/