修士論文

振幅包絡線情報の局部時間反転による音声プライバシー保護の研究

坂本 貴望

主指導教員 鵜木 祐史
審査委員主査 鵜木 祐史
審査委員 赤木 正人
党 建武
吉高 淳夫

北陸先端科学技術大学院大学
先端科学技術研究科
（情報科学）

令和３年２月

Abstract

Speech is one of the most important tools of communication in our social life. Speech contains three types of information: linguistic information, non-linguistic information, and para-linguistic information. Linguistic information is a verbal message that can be expressed in language and written. Non-linguistic information relates to the speaker's age, gender, personality, emotions, and so on. In general, these information cannot be intentionally controlled by the speaker. Para-linguistic information includes accent, intonation, speaking speed, and voice pitch, all of which cannot be expressed in writing. These information are intentionally added by the speaker to transform the linguistic information. In particular, linguistic information, which includes the semantic content of a message, is important in speech communication. It is not a problem for people who are communicating with each other to talk about linguistic information. However, there are situations in which it is undesirable for linguistic information to be overheard by unintended listeners, such as when talk about private conversations. In situations where confidentiality is required, linguistic information needs to be protected strictly and appropriately.

In open spaces such as hospitals, pharmacies, and conference rooms, people often talk about personal or confidential information. If such private conversations are overheard by unintended listeners, personal or confidential information may be leaked. Problems arising from private conversations must be solved to protect speech privacy.

Akagi and Irie's research on speech privacy protection focuses on mishearing of speech. This method is based on Acceleration of perceptual fusion. Acceleration of perceptual fusion is a phenomenon in which two or more different sounds are perceived as a single sound. It is occured according to Bregman's psychoacoustical heuristic regularities. This regularities composed of the following four rules: (1) common onset or offset, (2) smoothness, (3) harmonicity, (4) common changes occurring in the acoustic events. In this research, the target speech (conversational speech) and sound for hearing protection are presented simultaneously. It is shown that the two sounds are heard as one, and the speech content is obscured. While this method is very effective as a method of speech privacy protection, it has the problem of causing discomfort due to excessive deformation of the spectral envelope.

In auditory perception, the temporal structure of speech sounds has a significant effect on listening comprehension. For example, in Drullman's research, it is investigated whether the temporal amplitude envelope(TAE) or the temporal fine structure(TFS) of a sound contains more information related to speech intelligibility. The results showed that TAE plays an important role. In addition,

in Ueda's research is showed that speech intelligibility is significantly reduced by locally time-reversing. Locally time-reversing is a method of dividing speech into short segments, reversing the time within each segment, and then connecting the segments. It has been shown that when the locally time-reversal length is short (20-40 ms), the speech can be understood, but as the locally time-reversal length increases, the speech cannot be understood.

These results suggest that an effective method of speech privacy protection can be achieved by locally time-reversing TAE. This paper aims to achieve an effective method of speech privacy protection by directly processing spoken language information in time domain. In this paper, the following three points are focused on. (1) TAE plays an important role in speech intelligibility. (2) target speech and locally time-reversed speech have the same TFS based on the most restrictive of Bregman's four psychoacoustical heuristic regularities, rule(3): harmonicity. (3) locally time-reversing is used to manipulate the temporal structure of speech. From previous study, it has been found that two speech with different TAE (one of them locally time-reversed TAE) and the same TFS are most perceptually fused than those with different conditions from this. Therefore, in this paper, it is investigated whether locally time-reversing TAE can be perceptually fused to the target speech and reducing the speech intelligibility of the target speech under similar conditions.

In this study, two experiments were conducted. The first experiment was conducted in a soundproof room using headphones. The second experiment was conducted in an audio-visual laboratory using speakers. In the first experiment, it is considered whether or not that intelligibility of a target speech can be reduced with locally time-reversing TAE while TFS of the target speech has not been manipulated. Word intelligibility test with familiarity-controlled word lists was conducted to clarify these two points under two conditions of the highest and lowest word-familiarity levels and under nine conditions of locally time-reversing length (20, 40, 80, 160, 240, 320, 480, 640 ms, and whole duration). Results were summarized as: (1) the word recognition rate is reduced by controlling locally time-reversal length. (2) reduction degree of the word recognition rate depends on word-familiarity levels, that is, 77% at the highest familiarity level and 44% at the lowest familiarity level. (3) the most effective length of locally time-reversal is around 160 ms. These findings indicate that locally time-reversing TAE enables to reduce speech intelligibility of the target speech.

In the second experiment, the effectiveness of the method in a real environment is confirmed. In this test, the conditions of word-familiarity levels and locally time-reversing length were the same as in the previous test, and the target speech and locally time-reversing speech were played from different speakers. The results

showed that locally time-reversing TAE enables to reduce speech intelligibility of the target speech in a real environment It is also found that the effect is greater than that of the previous test that using headphones.

In both experiments, it was possible to reduce speech intelligibility of the target speech. These results suggest that locally time-reversing TAE enables to reduce speech intelligibility of the target speech. Also, it is shown that this method can effectively protect speech privacy.

目 次

図 目 次

第1章　序論

1.1　はじめに

我々が社会的生活を営む上で，音声は重要なコミュニケーション手段の一つである．まず，話し手が聞き手に伝えたいメッセージを頭に思い浮かべて，言語化する．そして，話し手はそのメッセージに基づき音声器官を動かして，音声を発する．その音声が聞き手の聴覚器官を介して脳に届き，メッセージが頭の中で理解される．音声コミュニケーションは，このような「ことばの鎖（Speech Chain）」と呼ばれる連鎖によって実現されている [1]．音声コミュニケーションでは，話し手が伝えようとしたメッセージはもちろん，その他にもさまざまな情報が伝達される．例えば，話し手の個人性や感情，意図などが挙げられる [2,3]．このように，音声に含まれる多様な情報を伝達しあうことで，円滑なコミュニケーションが成り立っている．

音声には大別して，言語情報，非言語情報，パラ言語情報の 3 種類の情報が含まれている [4]．言語情報とは，言語で表現できる，あるいは文脈から推測できる情報のことである．すなわち,「何を話しているか」という文字で書き表すことができる言語メッセージのことである．非言語情報とは，話者の年齢や性別，個人性，感情など身体・精神状態に関する情報のことである．これらの情報は一般的に，話し手が意図的に制御できない情報である．パラ言語情報とは，文字では表せないアクセント，イントネーション，話す速さや声の高さなどの情報のことである．これらの情報は，言語情報を変形させるため，または補完するために話者によって意図的に付加される情報である．

音声に含まれる情報の中でも特に，メッセージの意味内容が含まれる言語情報は，音声コミュニケーションにおいて欠かせない情報である．コミュニケーションを行っている当事者同士で意見を交換し，言語情報を伝達し合うことは問題にはならない．しかし，個人情報や機密情報に係る内容が話題に上る場面など，第三者に言語情報が伝達されることが望ましくない場合もある．このような秘話性が必要とされる場面では，言語情報は厳重かつ適切に保護される必要がある．

1.2 研究背景

病院や薬局，会議室といった特定のオープンスペースでなされる会話では，個人情報や機密情報に係わる内容が話題に上ることが多い．このようなプライバシーに係わる情報が会話とは関係のない第三者に聴き取られてしまった場合，個人情報や機密情報の漏洩につながってしまう．そのため，音声会話に含まれるプライバシーの漏洩防止，あるいはプライバシー保護の確立は喫緊の課題である [5]．音声会話に含まれるプライバシー，すなわち音声プライバシー（Speech Privacy）は，会話が当事者以外の人物に伝わらない状態（Confidential Privacy），あるいは会話が他者の執務を妨げない状態（Normal Privacy）として定義される [6]．音声プライバシーを適切に保護するためには，この二つの状態を常に保ち続けなければならない．

会話が第三者に聴き取られないような状態を保つためには，目的とする会話音声を不明瞭化（音声了解度を低下させること）させる必要がある．すなわち，目的音声が持つ音声言語情報が第三者に伝達されない状況を作り出す必要がある．音声言語情報を処理することで音声プライバシーを保護する方法として，室の音環境を変化させる方法と，目的音声そのものが持つ情報を操作する方法の二つがある．本研究では，前者を音声言語情報を間接的に処理する方法，後者を直接的に処理する方法と定義する．

音声言語情報を間接的に処理する方法として，音声伝送指標（STI: Speech Transmission Index）[7] を基準とした方法がある [8]．この方法は，STI のブラインド推定 [9,10] を利用して室における音声の聴き取りの状況を考慮し，目的音声に対して適切に残響を付与する方法である．この方法では，室内インパルス応答（RIR: Room Impulse Response）の後部残響を変化させ，STI を能動的に制御することで聴き取りにくさを制御し，音声プライバシー保護として非常に有効な方法の一つであることが示された．しかしながら，STI を低下させすぎた場合には，わずらわしさが増加してしまうという問題が残されている．

音声言語情報を直接的に処理する方法として，音情景解析の概念に基づく方法がある [11]．この方法は，音声知覚上で異聴を招くことを狙いしたものであり，「知覚的融合」に着目した方法である．この方法では，会話音声と同時に音声の音韻性を曖昧にする防聴音を提示し，二つの音が知覚的に混ざり合って一つの音として聴取されることで発話内容が不明瞭になることが示された．しかし，過度なスペクトル包絡の変形により知覚的に融合された音に不快感を生じさせるという問題を抱えている．

聴知覚において，音の時間構造は音の聴き取りに大きな影響を与える．例えば，時間情報による音声言語知覚に関する Drullman の研究がある [12]．この研究では，音の時間的振幅包絡線情報と時間微細構造のどちらに音声了解度に係わる情報が多く含まれているかを調査した．その結果，音の時間的振幅包絡線情報が重要な役割を担うことがわかった．また，音声を局部時間反転し，音声の時間構造を壊

すことによって音声了解度が著しく低下することも報告されている [13–15].

筆者はこのような背景を俯瞰的に眺めたところ，時間領域において言語情報を直接的（時間構造の操作）かつ間接的（残響付与）に処理したものを，Bregman の発見的規則に基づいて知覚的融合を促進するような形で目的音声に付与することで，効果的な音声プライバシー保護の方法を実現できるのではないかという考えに至った．二つのアプローチを組み合わせることで，目的音声の聴き取りにくさを担保しながら，STI を低下させすぎることによって生じるわずらわしさを低減させられることが見込める．本研究では，新たな音声プライバシー保護の方法を実現するための第一歩として，時間領域において音声言語情報を直接的に処理する方法を提案する．

1.3 研究目的

本研究の目的は，時間領域において音声言語情報を直接的に処理し，効果的な音声プライバシー保護の方法を実現することである．そこで，音声の振幅包絡線情報を局部時間反転することで音声言語情報を処理した音声を，知覚的融合を促進するような形で目的音声に付与することで目的音声を不明瞭化できるかを検討する．

まず，音声の時間構造の操作によって目的音声を不明瞭化できるかを明らかにするために，目的音声の振幅包絡線情報のみを局部時間反転した音声を用いて，音声の不明瞭化を検討する．

次に，本手法を用いた音声プライバシー保護の方法を実現するための第一歩として，実環境下での本手法の有効性を確認する．

最後に，振幅包絡線情報を局部時間反転した音声を用いた，音声プライバシー保護の方法を実現することが可能であるかについて考察する．

1.4 論文構成

本論文は，7 章で構成される．

第 1 章

音声コミュニケーションにおける言語情報の重要性について述べた後，音声プライバシー保護の研究背景と問題点を述べる．さらに，音の聴き取りに係る関連研究を説明した上で，本研究の目的を明らかにする．

第 2 章

音声プライバシー保護に関する研究に関して述べた後，音声言語知覚や音の聴き取りに関する研究に関して述べる．

第 3 章

本研究では，音声の振幅包絡線情報を局部時間反転することで，音声の不明瞭化を検討する．第 3 章では，はじめに本研究で想定する音声プライバシー保護の場面を説明し，次に局部時間反転処理について説明する．最後に，本研究の着目点について述べる．

第 4 章

目的音声の振幅包絡線情報のみを局部時間反転した音声を用いて，音声の不明瞭化を検討する．第 4 章では，目的音声とその振幅包絡線情報のみを局部時間反転した音声を加算して再生し，単語了解度試験を行う．

第 5 章

本手法を用いた音声プライバシー保護の方法を実現するための第一歩として，実環境下での本手法の有効性を確認する．第 5 章では，実環境を想定し，スピーカから音声刺激を空中放射して単語了解度試験を行う．

第 6 章

第 4 章，第 5 章で得られた結果から，振幅包絡線情報を局部時間反転した音声を用いた，音声プライバシー保護の方法を実現することが可能であるかについて全体考察を論じる．

第 7 章

本研究で明らかにしたことについて要約し，残された課題について述べる．

第2章　関連研究

2.1　音声プライバシー保護に関する研究

これまでに，音声プライバシー保護に関して，さまざまな観点から研究がなされている．物理的に会話音声の漏洩を防ぐ方法として，遮蔽効果の高いパーティションを設置し，話者を隔離する方法がある [16]．この方法は，室内空間の遮蔽が許される場合は最も有効である．しかし，オープンスペースのように室内空間の遮蔽が許されない場合には利用できないため，適用範囲が限定的であるという問題を抱える．

オープンスペース等での利用を考慮した方法として，マスキング法がある．この方法では，目的音声をマスクする別の音声を提示して目的音声のプライバシーを保護する．この方法の背後にある考えは，聴覚マスキングや情報マスキングに基づくものである．提示する音として，ピンク雑音や重畳雑音（バブル雑音），背景音楽などが一般的に利用される．中でも，帯域制限したピンク雑音が最も有効であるとされている [17, 18]．この方法は，目的音声そのものを聴き取りにくくすることで音声プライバシー保護を実現できる一方で，マスキング音を常時提示しておく必要があり，騒音暴露という根本的な問題を抱える．

騒音暴露を考慮した別のアプローチとして，残響特性を重畳することで音声プライバシー保護を実現する方法がある [19]．この方法は，会話音声に RIR を畳み込んで作成した残響音声を第三者に提示することで，目的音の聴き取りの妨害を狙うものである．この方法の背後にある考えは，音声の明瞭度・了解度は残響の影響により著しく低下するという知見に基づくものである．音声プライバシー保護を実現するためには，第三者の聴取環境に合わせて適切な残響付与が望まれるが，その方法の実現には至っていない．

この問題点に対し，STI に基づき，室における音声の聴き取りの状況を考慮した音声プライバシー保護の方法がある [8]．STI は，室の音声伝送品質の評価に利用される客観評価尺度であり，主観評価尺度である聴き取りにくさと高い相関を持つ指標である [7]．STI は，変調伝達関数（MTF: Modulation Transfer Function）の概念に基づいており，RIR から算出される．RIR は，直接音，初期反射，後部残響の三つの成分で構成され，後部残響が最も音声の聴き取りに影響を与えるとされている．この方法の背後にある考えは，STI が聴き取りにくさと高い相関を持つことと，RIR の後部残響が最も音声の聴き取りに影響を与えることに基づくものである．この方法では，直接音と後部残響で構成した RIR を用いて，そのパラ

メータを操作しSTIを能動的に制御することで，聴き取りにくさを制御した．この方法は，音声プライバシー保護として有効な方法の一つと考えられるが，STIを低下させすぎると，わずらわしさの増加を招くという問題を抱える．

他方で，雑音や残響を物理的に付与する方法ではなく，音声知覚上で異聴を招くことを狙いとした，音情景解析の概念に基づく音声プライバシー保護の方法がある [11]．この方法は，会話音声と同時に，音韻性を曖昧にする防聴音が提示され，これら二つの音が知覚的に混ざり合って一つの音に聴取されることで，発話内容が不明瞭になることを狙った方法である．その背後にある考えは，Bregmanの聴覚情景解析で述べられた発見的規則 [20] に基づき,「知覚的融合」を促進させるものである．知覚的融合とは，ある条件のときに二つの音が一つの音として知覚される現象のことである．Bregmanの発見的規則は,（1）共通の立ち上がり/立ち下がりの規則,（2）漸近的変化に関する規則,（3）調波関係に関する規則,（4）共通運命の原理に関する規則という四つの規則で構成される．これらの規則は，一つの音源から一つの音が出ていると考えたときに，音が時間領域および周波数領域において持つ性質について表されたものである．図2.1に，四つの発見的規則の概要を示す．発見的規則すべてに従っている音は，一つの音脈として知覚される．図2.1（1）では，音の立ち上がり/立ち下がりが一致しない特徴が存在する．このとき，共通の立ち上がり/立ち下がりの規則に従わないため，二つの音脈が知覚される．図2.1（2）では振幅が急激に変化している特徴が存在する．このとき，変化の滑らかさが保たれず漸近的変化に関する規則に従わないため，二つの音脈が知覚される．図2.1（3）では倍音構造から外れている特徴が存在する．このとき，調波関係に関する規則に従わないため，二つの音脈が知覚される．図2.1（4）では音の時間的な振幅包絡線が類似しない特徴が存在する．このとき，共通運命の原理に関する規則に従わないため，二つの音脈が知覚される．このように，規則に従わない特徴が存在すると，二つの音脈が知覚される．二つの音脈が混在していても，ある一方の音脈のみを選択的に聴取できるような現象は,「カクテルパーティ効果」と呼ばれる [21]．カクテルパーティ効果が生じる要因として，音の始まりの違い，音の到来方向の違い，音の高さの違い，音色の違いの他に，言語に関する知識や経験など多くの要素が挙げられる [22]．人が音を聞く場合，まず音の始まりの違い，音の到来方向の違い，音の高さの違い，音色の違いなどを用いて音響的な特徴をバラバラに分け（分離），言語に関する知識や経験などを用いて似た者同士をグルーピングし（群化），繋がりよく並べることで音のまとまりを形成する（音脈形成）という三つの処理が行われている [23]．この分離，群化，音脈形成の一連の働きを分凝と呼ぶ．分凝の際には，Bregmanの発見的規則を適用することで，同じ性質を持つ音同士をまとめて一つの音脈として知覚することができると言える．音情景解析の概念に基づく方法は，聴取者が会話の意味を取り違えることを狙った方法であり非常に有効である反面，過度なスペクトル包絡の変形による不快感を生じさせるという問題を抱えている．

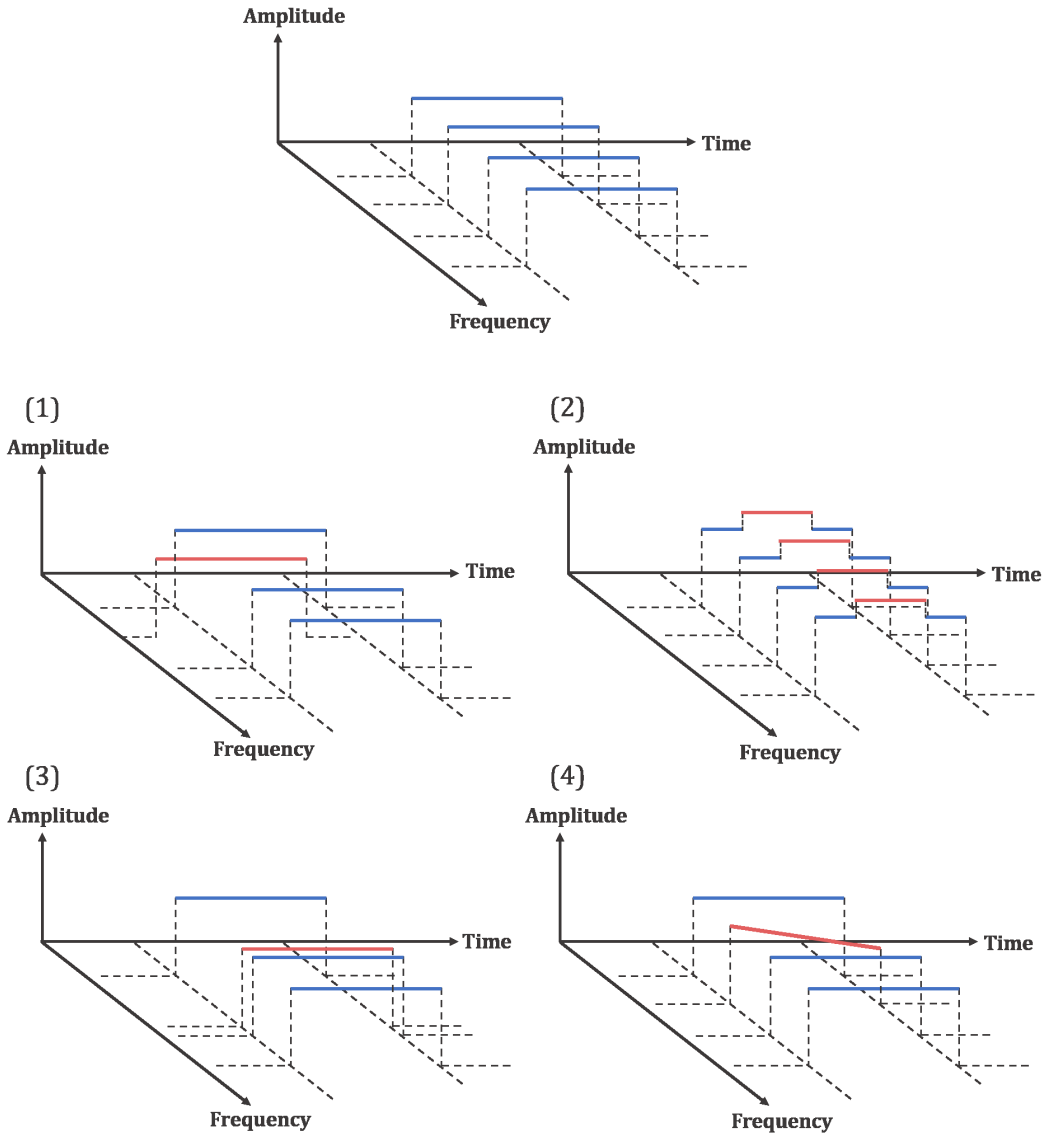

図 2.1: Bregman の発見的規則

2.2 音声言語知覚に関する研究

音声の振幅包絡線情報は，音声言語知覚において重要な役割を担っている．Drullman は，聴知覚メカニズムに基づき，音の時間的振幅包絡線情報と時間微細構造のどちらに音声言語知覚，特に音声了解度に係わる情報が多く含まれているかを調査した [12]．その結果，音の時間的振幅包絡線情報が重要な役割を担うことがわかった．また，Shannon らは帯域数を変化させた雑音駆動音声を用いて，振幅包絡線情報が音声言語知覚に与える影響を検討した [24–28]．雑音駆動音声とは，時間微細構造を白色性ガウス雑音に置き換え，音声の振幅包絡線情報のみを保存した音声のことである．その結果，音声を少なくとも 4 つの帯域で分割すれば, 振幅包絡線情報のみで音声言語知覚が可能であることが明らかになった．

音声の時間構造を壊すことは，音声言語知覚に大きな影響を及ぼす．時間構造の操作の一例として，時間反転が挙げられる．時間反転とは，音声を後ろから前へと逆再生する処理のことである．これまでに，発話音声を全区間で時間反転すると，人は発話内容を理解できなくなることがわかっている [29]．また，発話音声の時間反転によって，スペクトル形状を変化させずに自然性を保ったまま発話内容を理解できなくさせることがわかっている [30]．Ueda らは，音声を短い区間に区切り，それぞれの区間内で時間的に反転させる，つまり局部的に時間反転させることで音声了解度が著しく低下することを明らかにした [13]．この方法は局部時間反転と呼ばれ，反転区間長を変化させることで局所的時間反転音声の了解度を系統的に操作できることがわかっている．局部時間反転の反転区間長が 20〜40 ms と短い場合は発話内容を理解できるが，反転区間長がこれより長くなるにつれて発話内容を理解できなくなることが報告されている．

第3章　本研究の方略

3.1　問題設定

図 3.1 に，本研究で想定する音声プライバシー保護の場面を示す．オープンスペースにおいて，個人情報や機密情報に関する会話がなされている状況を想定する．ここでは会話音声 $x(t)$ が漏洩し，会話とは関係のない第三者が漏洩音声 $y(t)$ を聴き取るものとする．本研究では，$x(t)$ と同時に会話内容を不明瞭化させる音声 $x'(t)$ を第三者に漏洩させる．$x(t)$ と $x'(t)$ の二つの音を知覚的に融合させて，$x(t)$ が本来持つ言語情報とは異なる情報を持つ音声 $y(t) = x(t) + x'(t)$ として第三者に知覚させる．その結果，第三者が会話音声の意味を取り違え，音声の不明瞭化が見込める．

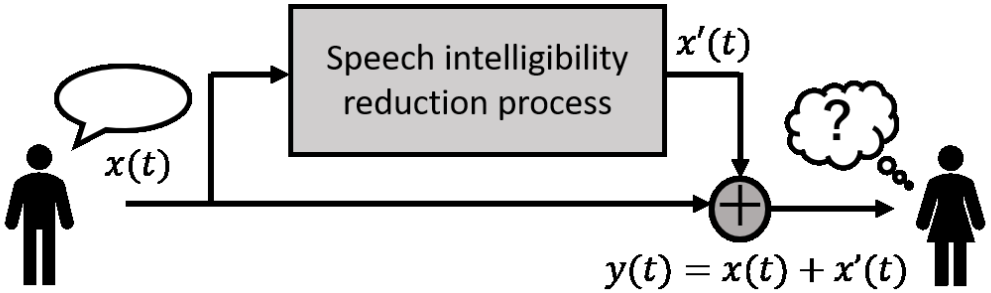

図 3.1: 問題設定：音声プライバシー保護の場面

3.2 局部時間反転

音声 $x(t)$ を不明瞭化させるために，時間構造の操作として局部時間反転処理を行う．局部時間反転とは，音声を短い区間に区切り，それぞれの区間内で時間的に反転させた後，再びつなぎ合わせることを指す．これまでに，発話音声を全区間で時間反転すると，人は発話内容を理解できなくなることがわかっている [29]. また，局部時間反転の反転区間長が 20〜40 ms と短い場合には発話内容を理解できるが，反転区間長がこれより長くなるにつれて発話内容を理解できなくなることが報告されている [13].

以上の報告を踏まえ，本研究では，音声信号そのものではなく，音声了解度に重要な役割を担う振幅包絡線情報の局部時間反転を行う．本研究では，時間微細構造は保存したまま，局部時間反転された振幅包絡線情報と組み合わせることで局部時間反転音声 $x'(t)$ を作成する．このとき，$x'(t)$ は，会話音声 $x(t)$ と同じ時間微細構造を持つ．そのため，Bregman の発見的規則（調波性に関する規則）に基づけば $x(t)$ と $x'(t)$ の二つの音が知覚的に融合され，漏洩音声 $y(t)$ として第三者に聴き取られる．そして，第三者が結局のところ， $y(t)$ から本来の発話内容を理解できなくなる状況が生み出される．

本研究では，時間微細構造を保存したまま振幅包絡線情報のみを局部時間反転して合成された音声のことを，以後,「局部時間反転音声」と呼ぶことにする．

3.3 局部時間反転音声の作成手順

本研究では，図 3.2 に示すような音声分析合成系を使用し，局部時間反転音声を作成する．この音声の周波数分解には，定帯域ガンマトーンフィルタバンクを使用する．ガンマトーンフィルタは聴覚フィルタ [31] の一つであり，健聴者の平均的な周波数分析特性の一次近似として提案されたものである [32]．ガンマトーンフィルタのインパルス応答は，以下のように表される [33]．

$$g_k(t) = at^{(N-1)}\exp(-2\pi b\mathrm{ERB}_N(f_k)t)\cos(2\pi f_k t + \phi), \tag{3.1}$$

f_k は聴覚フィルタの中心周波数，a は振幅，t は時間，N は次数，ϕ は位相である．聴覚フィルタの帯域幅である等価矩形帯域幅 $\mathrm{ERB_N}(f_k)$ は，以下のように表される [31]．

$$\mathrm{ERB}_N = 24.7 \times \left(\frac{43.7 f_k}{1000} + 1\right), \tag{3.2}$$

音声を周波数分解する際，サンプリング周波数を 20 kHz，各フィルタの帯域幅を 100 Hz とした．

はじめに，入力信号（原音）を定帯域ガンマトーンフィルタバンクによって 100 帯域に分割する．次に，各帯域信号から Hilbert 変換を利用して振幅包絡線と時間微細構造を求める．次に，振幅包絡線を局部時間反転し，時間微細構造はそのままとする．この処理をすべての帯域信号に対して行った後，ガンマトーンフィルタバンクの逆変換によって総加算処理が行われ，局部時間反転音声を得る．

図 3.3 に局部時間反転処理の例を示す．ここでは，対象となる振幅包絡線を指定された反転区間長で等間隔に分割し，それぞれの区間内で時間反転を行った後，反転された区間を再び時間軸上でつなぎ合わせる．時間微細構造は何も操作されずそのまま保存される．最後に，局部時間反転された振幅包絡線に保存された時間微細構造を掛け合わせることで局部時間反転された帯域信号が得られる．

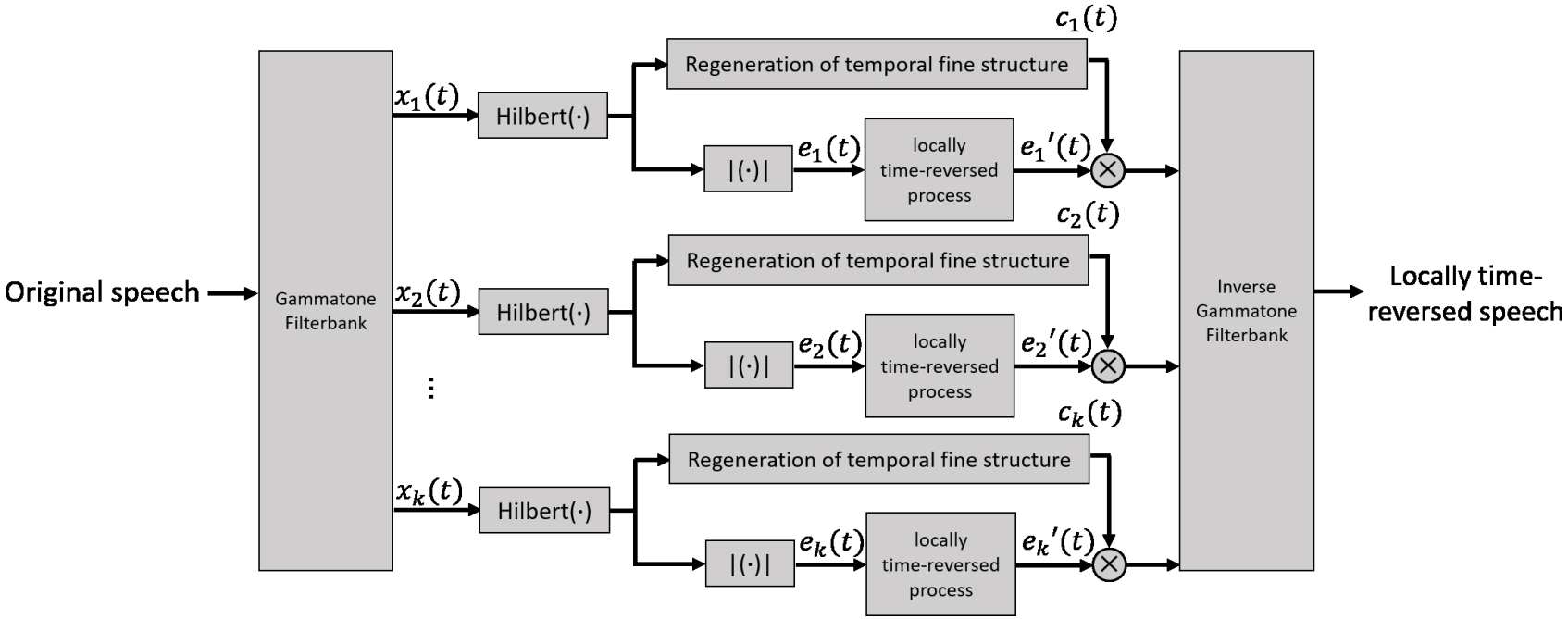

図 3.2: 音声分析合成系を利用した局部時間反転音声の作成手順

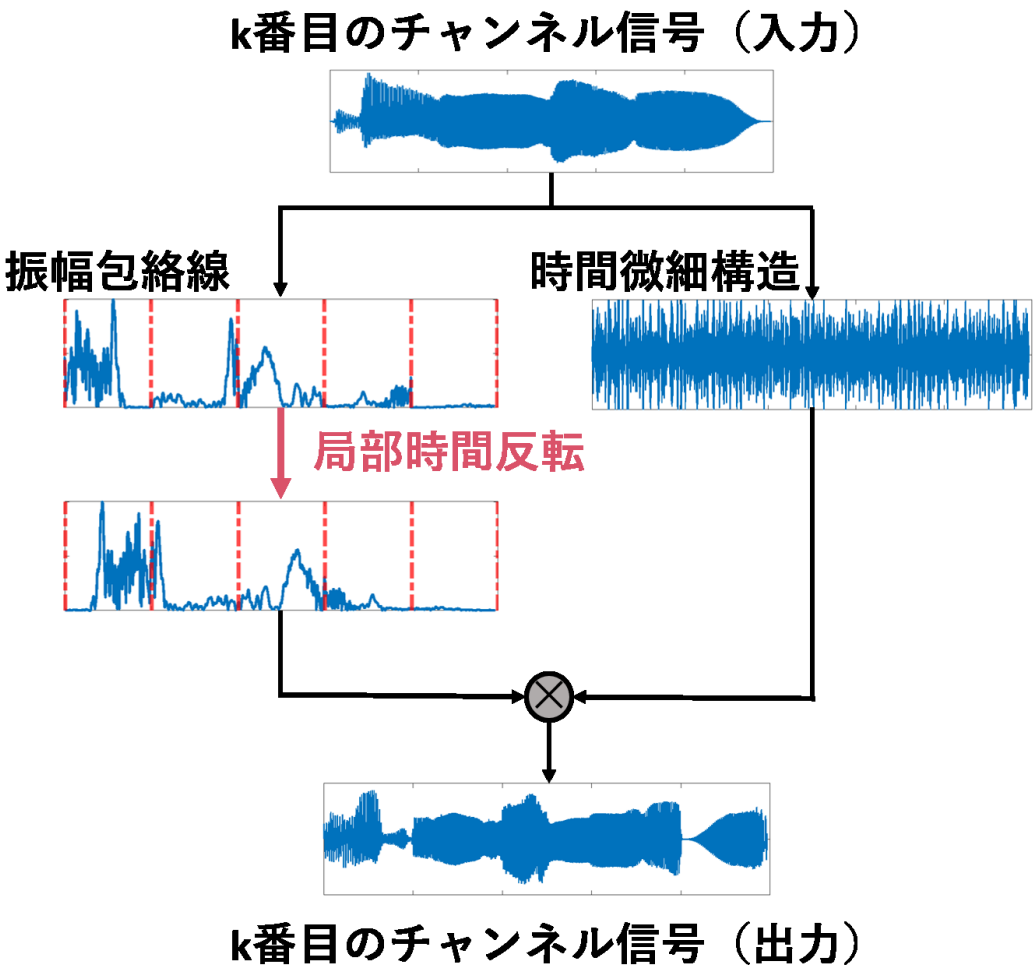

図 3.3: 各帯域信号における振幅包絡線の局部時間反転

3.4　本研究の着目点

音声の振幅包絡線情報は，音声言語知覚において重要な役割を担っている [12]．また，局部時間反転処理によって音声の時間構造を壊すことで，音声了解度は著しく低下する [13]．さらに，Bregman の発見的規則に基づき知覚的融合を促進させるような音声プライバシー保護の方法は，発話内容を不明瞭化できることが報告されている [11]．これらのことから筆者は，音声の時間構造を操作して音声言語情報を処理した音声を，知覚的融合を促進するような形で会話音声に付与することで，効果的な音声プライバシー保護の方法を実現できるのではないかという考えに至った．この考えを一つずつ確認するために，まず，音声の時間構造の操作と目的音声への知覚的融合の促進ならびに融合された音の不明瞭化を検討する必要がある．

そこで本研究では，次の 3 点に着目して，二つの音の融合と，融合された音の不明瞭化を検討する．(1) 音の時間的振幅包絡線情報が，音声了解度に重要な役割を担っていること，(2) Bregman の発見的規則の中でも最も制約の強い「調波性に関する規則」に基づき，提示音声と目的音声が同じ時間微細構造を持つこと，(3) 音声の時間構造の操作として局部時間反転処理を利用すること．本研究では，時間微細構造は保存したままで，振幅包絡線包絡線情報のみを局部時間反転した音声を目的音声に知覚的に融合させ，目的音声を不明瞭化できるかどうかを調査する．

本研究では音声の振幅包絡線情報を局部時間反転し，時間領域において目的音声が持つ音声言語情報を操作する．そのため，スペクトル包絡が保存され，音情景解析の概念に基づく方法で問題となっている不快感は生じないことが見込める．

第4章　振幅包絡線情報の局部時間反転による音声の不明瞭化

4.1　実験目的

本実験の目的は，目的音声とその振幅包絡線情報のみを局部時間反転した音声を加算して再生した場合に，目的音声の了解度が低下するかどうかを明らかにすることである．その際，最も効果的な反転区間長を明らかにするために，反転区間長を変化させて実験を行った．また，音声了解度は単語親密度によって大きく影響される．高親密度の単語においても本手法の効果があるかどうかを確かめるため，低親密度と高親密度の単語を用いて実験を行った．

4.2　実験方法

4.2.1　被験者

実験には，日本語を母語とし正常聴力を有する成人 10 名（23-27 歳，男性 6 名，女性 4 名）が参加した．

4.2.2　装置と刺激

実験刺激は，PC（LG Sharkoon, Windows8.1）より，A/D コンバータ（RME FIREFACE UCX），およびヘッドホンアンプ（audio-technical AT-HA5000）を経由して密閉型ヘッドホン（SENNHEISER HDA200）から被験者に提示した．被験者の反応の取得には MATLAB にて作成した GUI アプリケーションを使用し，入力装置にはキーボード（ELECOM, TK-FCM084）を使用した．

音声刺激として，親密度別単語了解度試験用データベース（FW07）[34] の男女各 2 名の話者の 4 モーラ単語と，これらを原音として作成した局部時間反転音声を提示した．局部時間反転の反転区間長は 20, 40, 80, 160, 240, 320, 480, 640 ms, ALL（全区間）の 9 条件とした．単語親密度は 1.0〜2.5，5.5〜7.0 の 2 条件とした．刺激の総数は，360 個（=9 反転区間長条件× 2 親密度条件× 20 音声）であった．音声刺激は，人工耳（BK Artificial Ear Type 4153），マイク（BK Microphone

Type 4192），騒音計（BK Sound Level Meter Type 2250）を用いて，A 特性音圧レベルがおよそ 62 dB となるよう設定した．

4.2.3　手続き

単語了解度試験は防音室で行われた．実験環境の概略を図 4.1 に示す．被験者には，目的音声の 4 モーラ単語に局部時間反転音声を加算して提示した．被験者の課題は，聴き取った単語をカタカナでキーボード入力することであった．各被験者内で反転区間長および単語親密度条件はランダムな順で提示し，また話者 4 名の発話が均等に提示されるように調整した．反転区間長と単語リストの組み合わせは，被験者ごとに変更した．実験には，休憩を含め 1 時間程度を要した．

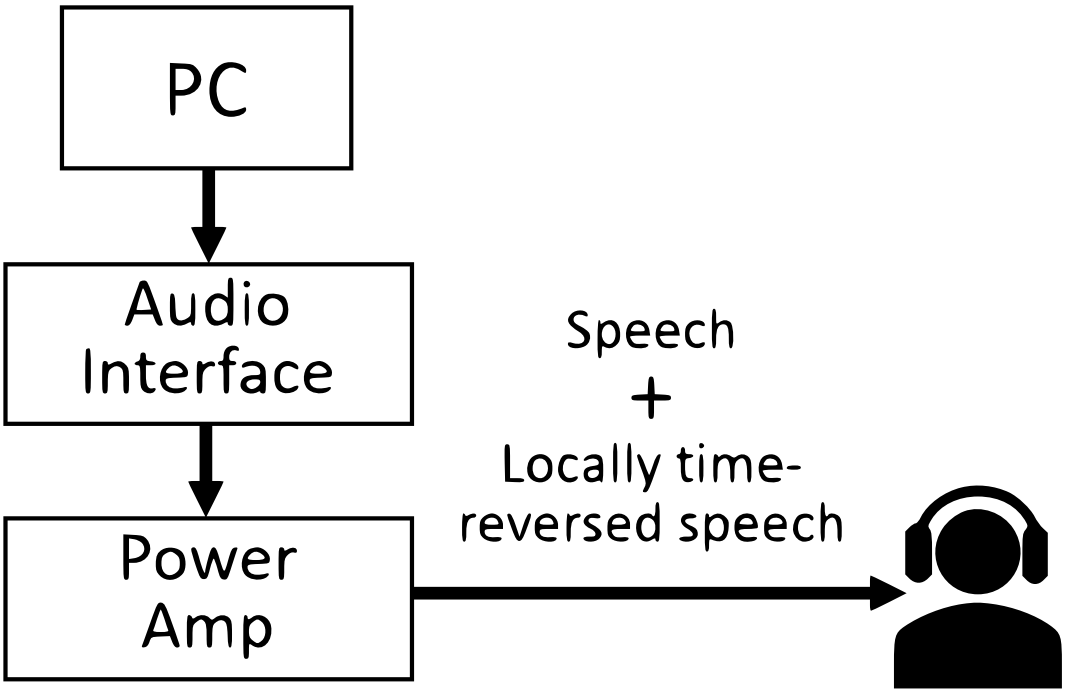
PC
Audio Interface
Power Amp
Speech
+
Locally time-reversed speech

図 4.1: 実験環境

4.3 実験結果

4.3.1 単語正答率

各被験者で条件ごとに単語正答率を算出した．図 4.2 に，親密度別単語了解度試験の結果を示す．

図中に，被験者 10 名の平均単語正答率と標準誤差を示す．図 4.2 から，両親密度条件ともに，反転区間長に応じて単語正答率が低下する傾向が見られた．ただし，単語親密度によって単語正答率の低下の度合いは異なった．低親密度の場合には最大で 44%まで単語正答率が低下した（ALL 条件）．一方，高親密度の場合には最低でも 77%程度にとどまった．

反転区間長条件と単語親密度条件を要因とした，2 要因分散分析を行った結果，反転区間長条件の主効果（$F(1, 162) = 250.16, p < 0.05$），および単語親密度条件の主効果（$F(8, 162) = 16.53, p < 0.05$）が認められた．さらに，反転区間長条件と単語親密度条件の交互作用が認められた（$F(8, 162) = 2.76, p < 0.05$）．交互作用が認められたことから，親密度別に反転区間長条件について多重比較（Bonferroni 法）を行った結果，低親密度の場合は反転区間長 20 ms と 80 ms〜ALL，40 ms と 160 ms〜ALL，80 ms と ALL に有意差が認められた ($p < 0.05$)．また，高親密度の場合は反転区間長 20 ms と 320 ms, 640 ms, ALL，40 ms と 160 ms〜ALL，80 ms と ALL に有意差が認められた ($p < 0.05$)．

以上の結果から，反転区間長に応じて単語正答率が低下することが示された．また，単語親密度によって反転区間長に対する単語正答率の低下の度合いが異なることが示された．

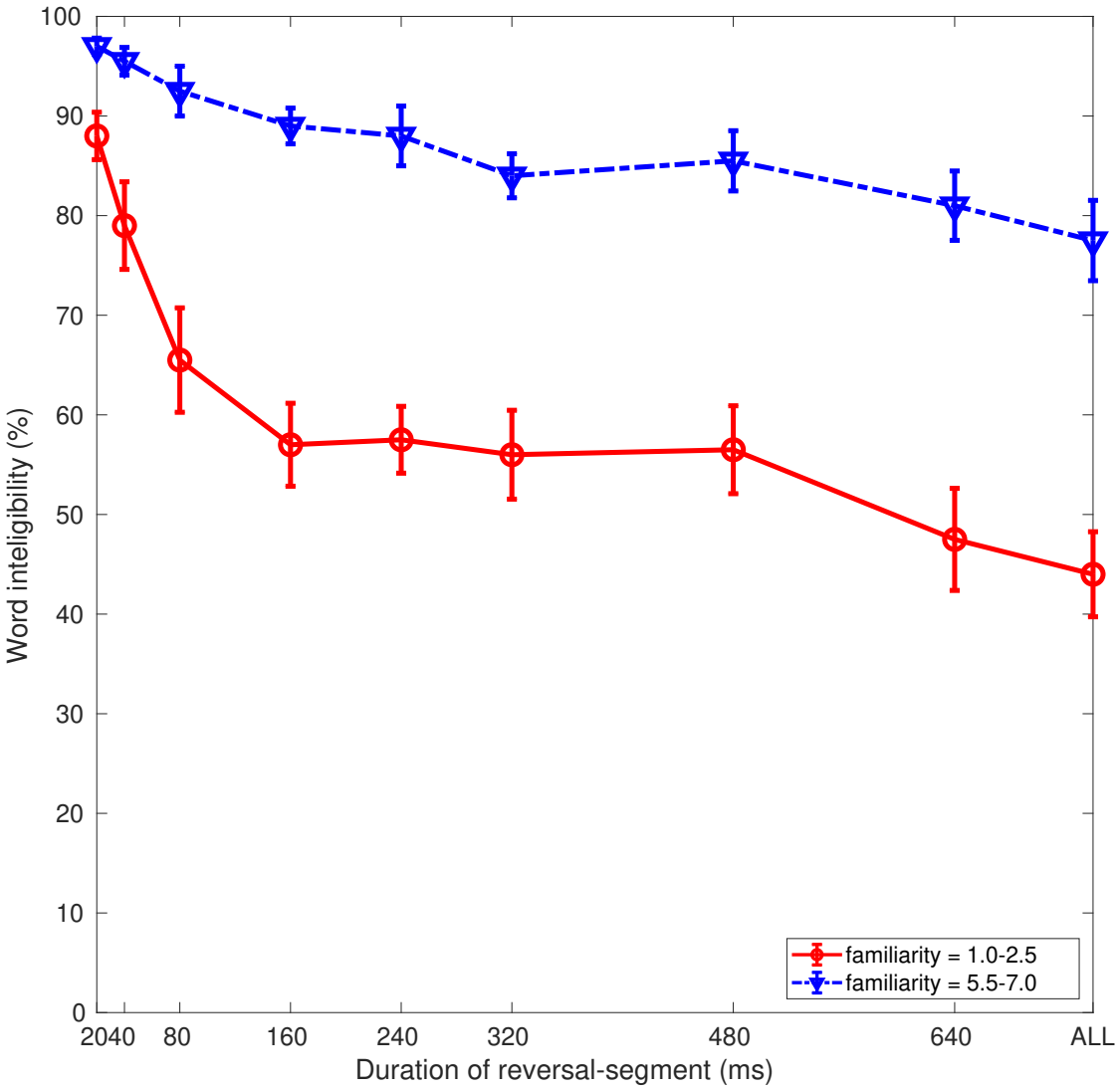

図 4.2: 親密度別単語了解度試験の結果

4.3.2 モーラ位置別正答率

モーラ位置によって正答率が変わるかどうかについて検討するため，各親密度で単語内のモーラ位置での正答率を算出した．図 4.3 に，低親密度単語のモーラ位置別の平均正答率と標準誤差を示す．反転区間長 20〜320 ms では第 1 モーラの正答率が最も高かった．また，反転区間長 160〜320 ms では各モーラ位置で正答率がほぼ変わらない傾向であった．

図 4.4 に，高親密度単語のモーラ位置別の平均正答率と標準誤差を示す．反転区間長 20〜240 ms では，第 1 モーラの正答率が最も高かった．また，反転区間長 160〜320 ms では各モーラ位置で正答率がほぼ変わらなかった．これは低親密度単語と同じ傾向であった．

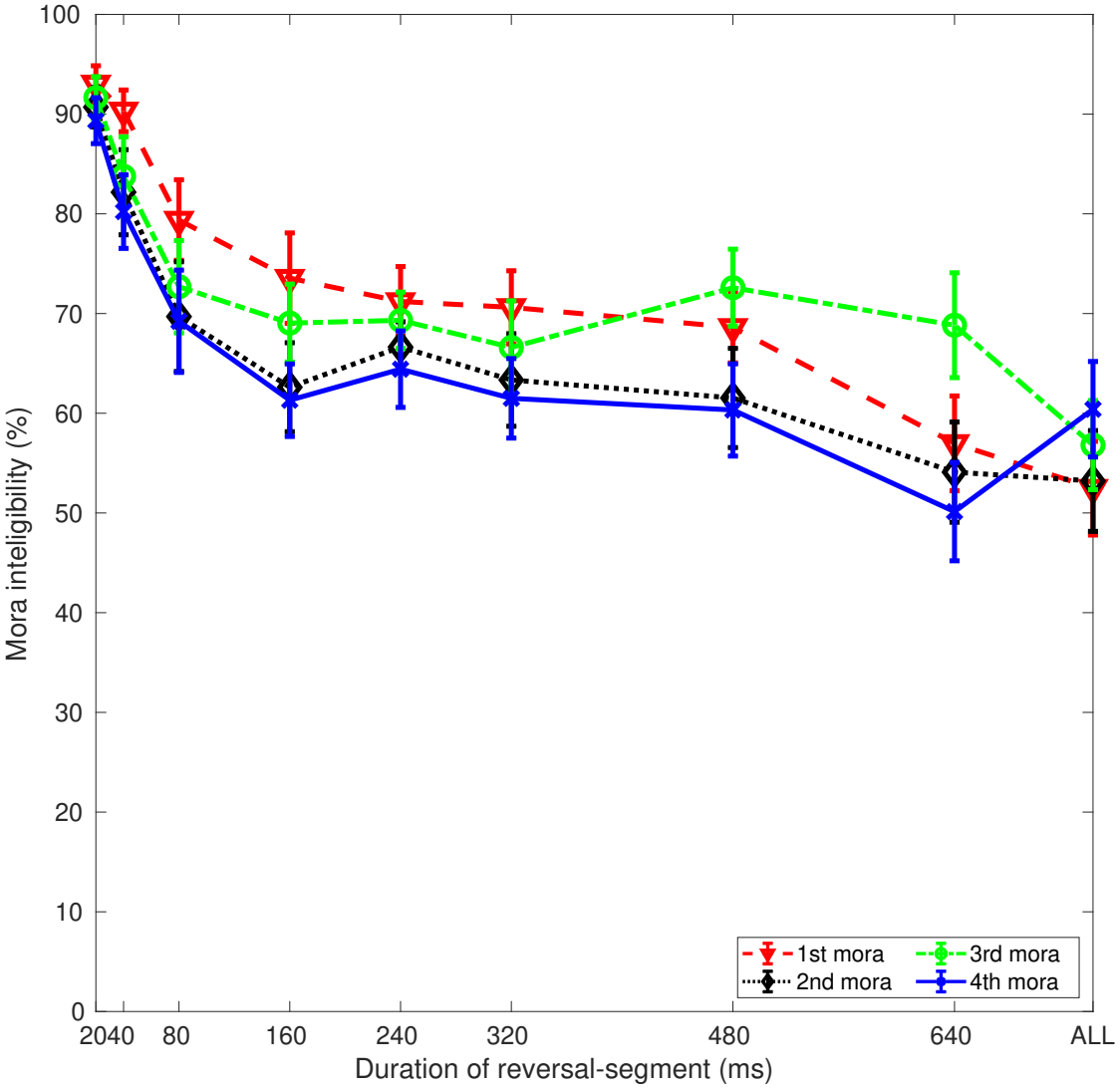

図 4.3: モーラ位置別正答率（親密度 1.0-2.5）

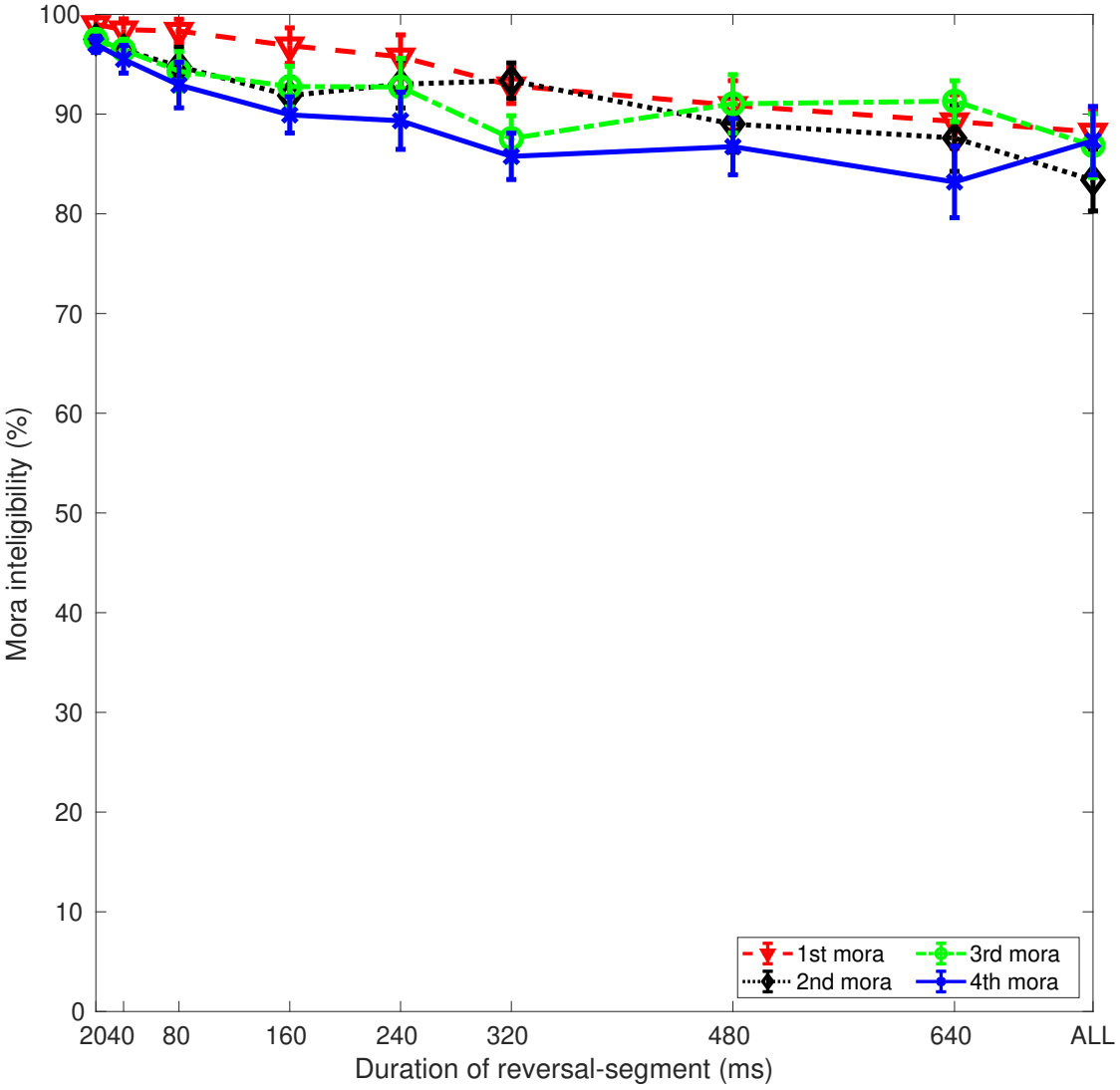

図 4.4: モーラ位置別正答率（親密度 5.5-7.0）

4.4 考察

まず，目的音声の振幅包絡線情報のみを局部時間反転した音声を用いて，目的音声を不明瞭化できるかについて議論する．本実験では，目的音声とその振幅包絡線情報のみを局部時間反転した音声の二音を加算して提示し，単語了解度試験を行った．その結果，反転区間長を操作することによって単語正答率が低下し, 最大で 44%まで成績が低下した．この結果から，音声の振幅包絡線情報を局部時間反転させただけで音声了解度を低下させられたことが確認できた．また，反転区間長に応じて目的音声の単語正答率が低下することから，音声の時間構造の操作として局部時間反転処理を利用できることを確認した．さらに，実験後に被験者の内観を聞いたところ，音声刺激が一つの音に聴こえたとの報告があった．このことから，目的音声と局部時間反転音声が同じ時間微細構造を持つことで，二音が知覚的に融合することが確認できた．これらのことから，振幅包絡線情報を局部時間反転した音声を用いることで，目的音声を不明瞭化できることが示された．

次に，本手法を用いて音声プライバシー保護を行う上で，有効な反転区間長について議論する．実験の結果，反転区間長に応じて単語正答率が低下し，ALL（全区間）で最も成績が低下した．しかし，単語親密度に関わらず，反転区間長 160 ms と ALL には有意差は認められなかった．このことから，反転区間長 160 ms では，全区間反転と同程度に音声を不明瞭化させられることが期待できる．実際のオープンスペースでの実装を考慮すると，目的音声を収録し反転して再生するため，反転区間長はできる限り短い必要がある．本研究で用いた音声刺激は，1 モーラあたりの時間長が約 180 ms であった．すなわち，160 ms の反転区間長は 1 モーラ未満であり，実装可能な区間長であると考えられる．

最後に，反転区間長 20〜160 ms では単語正答率が低下し，それ以降の反転区間長条件では成績が大きく低下しなかった理由について考察する．反転区間長が 160 ms 未満の条件（20〜80 ms）では，1 モーラに満たない反転区間長で時間反転した音声が目的音声に加算される．160 ms 未満の条件で単語正答率が高かったことは，各モーラの途中で時間反転した音声を目的音声と加算して提示しても，目的音声の各モーラを壊しきれないことを示す．一方，160 ms よりも長い反転区間長でも，同様に各モーラの途中で時間反転をする（例えば 240 ms は 1.3 モーラ前後，320 ms は 1.8 モーラ前後に相当する）．つまり，常に目的音声の第 1 モーラには各モーラの途中で時間反転した音声が加算される．このことから，160 ms よりも長い反転区間長では目的音声の第 1 モーラの正答率が低下しなかったと考えられる．一般に第 1 モーラが知覚できるとそれ以降のモーラも推定できるとされている [35]．反転区間長が 160 ms 以上の条件では第 1 モーラが知覚できたため，それ以降のモーラも推定でき，単語正答率が下げ止まったと考えている．

第5章　実環境下での音声の不明瞭化

5.1　実験目的

本実験の目的は，本手法を用いた場合に，実環境においても目的音声の了解度が低下するかを確認することである．第 4 章ではヘッドフォンを用いて，目的音声とその振幅包絡線情報のみを局部時間反転した音声を加算して再生することによって実験を行った．第 5 章では，実際の音声プライバシー保護の場面を想定して，目的音声と振幅包絡線情報をそれぞれ異なるスピーカから再生して実験を行い，本手法の有効性を確認する．

5.2　実験方法

5.2.1　被験者

実験には，日本語を母語とし正常聴力を有する成人 10 名（23-32 歳，男性 7 名，女性 3 名）が参加した．

5.2.2　装置と刺激

実験刺激は，PC（Windows10）より，A/D コンバータ（RME FIREFACE UCX），およびアンプ（YAMAHA, A-U671）を経由してスピーカ（YAMAHA, NS-pf7）から被験者に提示した．被験者の反応の取得にはタブレット PC（Surface Pro3）を使用した．

音声刺激として，親密度別単語了解度試験用データベース（FW07） の男女各 2 名の話者の 4 モーラ単語と，これらを原音として作成した局部時間反転音声を提示した．局部時間反転の反転区間長は 20, 40, 80, 160, 240, 320, 480, 640 ms, ALL（全区間）の 9 条件とした．単語親密度は 1.0〜2.5，5.5〜7.0 の 2 条件とした．刺激の総数は，360 個（=9 反転区間長条件× 2 親密度条件× 20 音声）であった．音声刺激は，各スピーカからピンク雑音を提示し，受聴位置において騒音計（RION, NL-42）にて A 特性音圧レベルがおよそ 62 dB となるよう調整した上で，目的音声の 4 モーラ単語と局部時間反転音声がおよそ 62 dB となるよう設定した．

5.2.3 手続き

単語了解度試験は AV 実験室で行われた．実験環境の概略を図 5.1 に示す．局部時間反転音声を提示するスピーカを被験者から 1.5 m の位置へ，目的音声の 4 モーラ単語を提示するスピーカを被験者から 3 m の位置へ設置して，被験者に音声刺激を提示した．被験者の課題は，聴き取った単語をカタカナでキーボード入力することであった．被験者には，必ずスピーカの方を見て音声を聴き取るように教示した．各被験者内で反転区間長および単語親密度条件はランダムな順で提示し，また話者 4 名の発話が均等に提示されるように調整した．反転区間長と単語リストの組み合わせは，被験者ごとに変更した．実験には，休憩を含め 1 時間程度を要した．

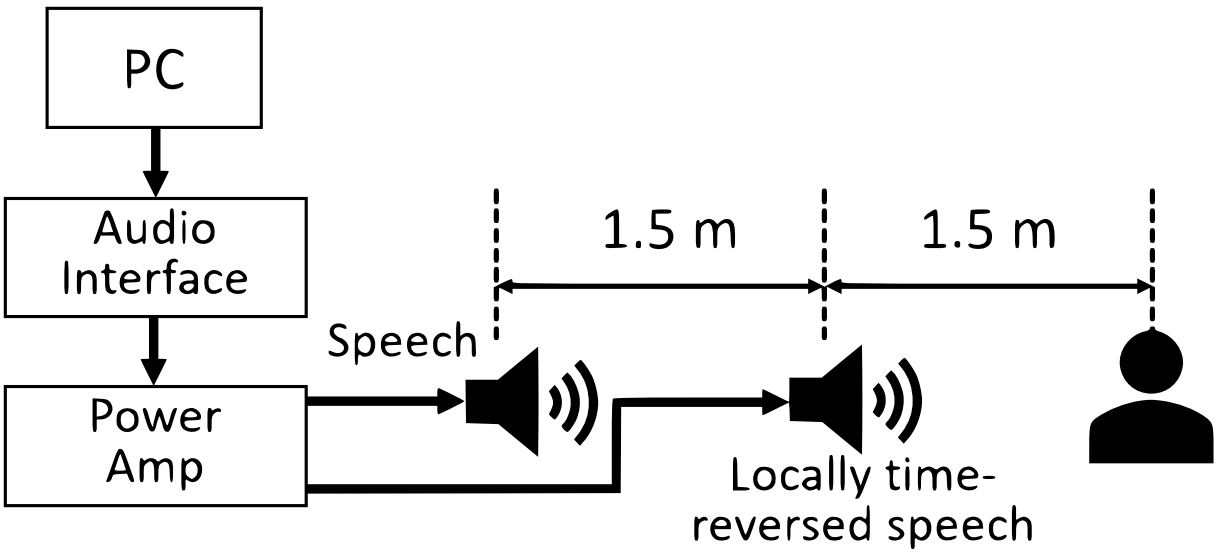
PC
Audio Interface
Power Amp
Speech
1.5 m
1.5 m
Locally time-reversed speech

図 5.1: 実験環境

5.3 実験結果

5.3.1 単語正答率

各被験者で条件ごとに単語正答率を算出した．図 5.2 に，親密度別単語了解度試験の結果を示す．

図中に，被験者 10 名の平均単語正答率と標準誤差を示す．図 5.2 から，両親密度条件ともに，反転区間長に応じて単語正答率が低下する傾向が見られた．ただし，単語親密度によって単語正答率の低下の度合いは異なった．低親密度の場合には最大で 20%まで（ALL 条件），高親密度の場合には最大で 48%まで（反転区間長 640 ms 条件）単語正答率が低下した．

反転区間長条件と単語親密度条件を要因とした，2 要因分散分析を行った結果，反転区間長条件の主効果（$F(1, 162) = 152.57, p < 0.05$），および単語親密度条件の主効果（$F(8, 162) = 44.60, p < 0.05$）が認められた．また，反転区間長条件と単語親密度条件の交互作用は認められなかった（$F(8, 162) = 1.10, p < 0.05$）．それぞれの親密度で反転区間長条件について多重比較（Bonferroni 法）を行った結果，低親密度の場合は反転区間長 20 ms と 80 ms〜ALL，40 ms と 160 ms〜ALL，80 ms と 240 ms〜ALL，160 ms と 640 ms〜ALL，240 ms と ALL に有意差が認められた ($p < 0.05$)．また，高親密度の場合は反転区間長 20 ms と 240 ms〜ALL，40 ms と 240 ms〜ALL，80 ms と 240 ms〜ALL，160 ms と 640 ms〜ALL に有意差が認められた ($p < 0.05$)．

以上の結果から，反転区間長に応じて単語正答率が低下することが示された．また，反転区間長に対する単語正答率の低下の度合いは単語親密度によらず同程度であることがわかった．

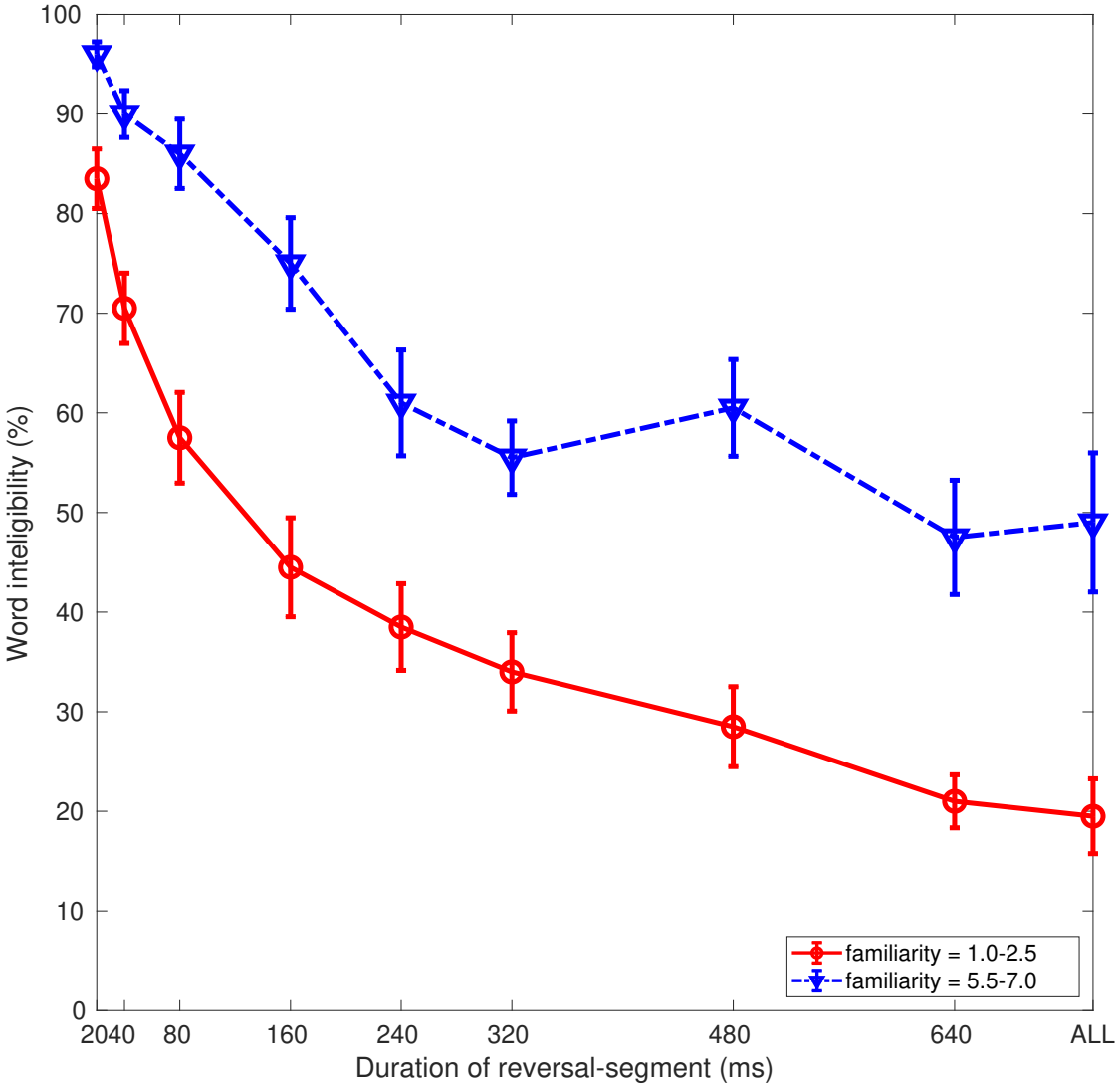

図 5.2: 親密度別単語了解度試験の結果

5.3.2　モーラ位置別正答率

各親密度で，単語内のモーラ位置での正答率を算出した．図 5.3 に，低親密度単語のモーラ位置別の平均正答率と標準誤差を示す．反転区間長 20〜160 ms では第 1 モーラの正答率が最も高かった．また，反転区間長 320〜640 ms では第 3 モーラの正答率が最も高かった．

図 5.4 に，高親密度単語のモーラ位置別の平均正答率と標準誤差を示す．反転区間長 20〜320 ms では，第 1 モーラの正答率が最も高かった．また，反転区間長 480〜640 ms では第 3 モーラの正答率が最も高かった．

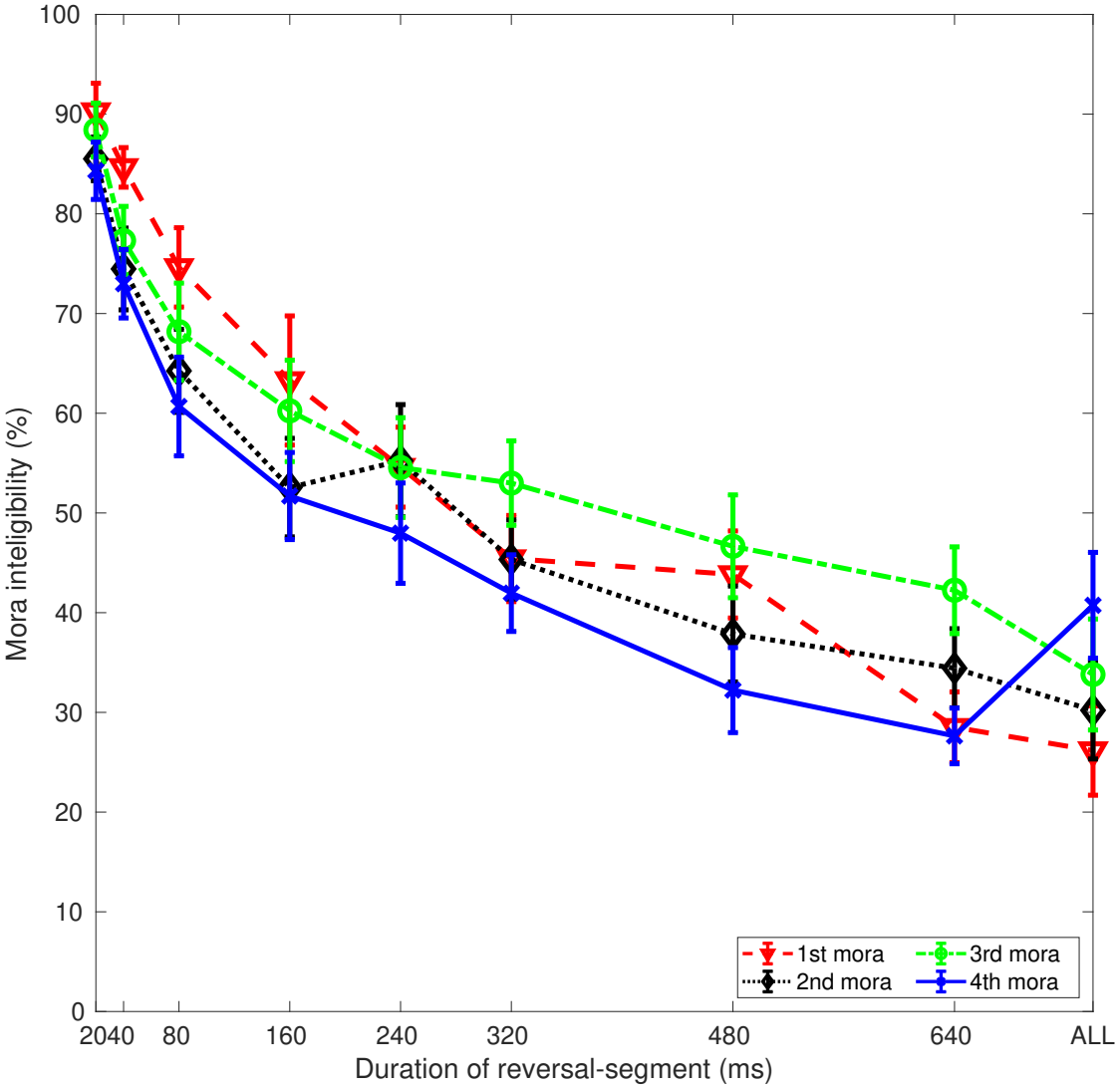

図 5.3: モーラ位置別正答率（親密度 1.0-2.5）

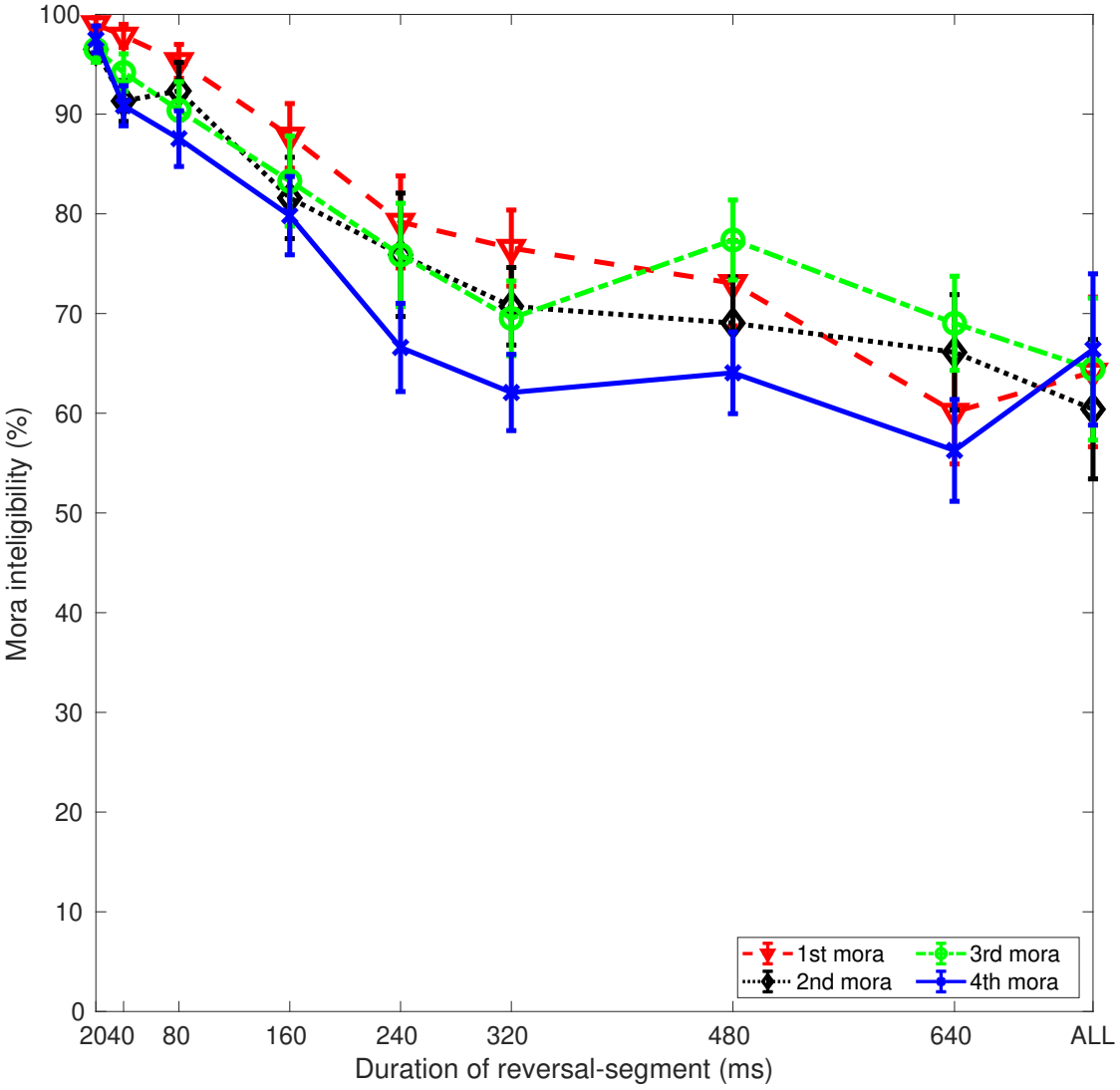

図 5.4: モーラ位置別正答率（親密度 5.5-7.0）

5.4 考察

目的音声の振幅包絡線情報のみを局部時間反転した音声を用いることで，実環境下で目的音声を不明瞭化できるかについて議論する．本実験では，目的音声とその振幅包絡線情報のみを局部時間反転した音声の二音をそれぞれ異なるスピーカから提示し，単語了解度試験を行った．その結果，反転区間長を操作することによって単語正答率が低下し, 最大で 20%まで成績が低下した．このことから，実環境下で，振幅包絡線情報の局部時間反転によって音声了解度を低下させられたことが確認できた．また，反転区間長に応じて目的音声の単語正答率が低下することから，実環境下で局部時間反転処理を利用できることが確認できた．さらに，実験後に被験者の内観を聞いたところ，音声刺激が何を言っているかわからない一つの音として聴こえたという報告があった．このことから，実環境下で目的音声と局部時間反転音声の二音が知覚的に融合することが確認できた．これらのことから，実環境において，振幅包絡線情報を局部時間反転した音声を用いることで目的音声を不明瞭化できることが示された．

第6章　全体考察

第 4 章では，目的音声の振幅包絡線情報のみを局部時間反転した音声を用いることで，目的音声を不明瞭化できるかを調査した．その結果，反転区間長を操作することによって音声を不明瞭化できることが明らかになった．また，反転区間長 160 ms では全区間反転と同程度まで了解度を下げられることが明らかになった．第 5 章では，実環境下での本手法の有効性を調査した．その結果，実環境においても反転区間長の操作によって音声を不明瞭化でき，本手法が有効であることが確認できた．以上の結果を関連付けると，音声の振幅包絡線情報を局部時間反転することで，目的とする音声を不明瞭化できることが明らかになった．

ヘッドフォンを用いて実験を行った場合と比較して，実環境を想定してスピーカを用いて実験を行った場合のほうが音声の不明瞭化が大きくなった理由について考察する．実環境を想定した実験は，AV 実験室にてスピーカから音声刺激を空中放射して行った．ヘッドフォンを用いて実験を行った場合とは異なり，実験室内には残響が存在している．残響とは，音が壁や天井等で繰り返し反射することで，音声が響いて聴こえる現象のことである [36]．室の残響特性は，TSP 信号を用いて測定される RIR から知ることができる [37]．すなわち，音源から発せられた音には RIR が重畳され，聴取者の耳に届くこととなる．そのため，残響のある室においては重畳成分の影響で音声の特徴が歪み，聴き取りにくくなる．本実験では，目的音声と局部時間反転音声の二つの音の残響成分が重なることで混合音として聴取され，知覚的に融合しやすくなったことが考えられる．そのため，ヘッドフォンを用いた実験と比較して，音声の不明瞭化が大きくなったと考えられる．また，実環境においては残響だけでなく雑音の影響も見逃すことはできない．音声信号に対する雑音の影響には，加法性雑音と乗法性雑音がある [38]．加法性雑音は，室における雑音のように，音声信号に対して雑音が加法的に影響する雑音のことである．乗法性雑音は，通信路における雑音のように，音声信号に対して雑音が乗法的に影響する雑音のことである．音環境においては音声に雑音が加法され，同時マスキングが起こる．また，雑音が存在する環境では，パワーが小さい子音などの成分はマスクされやすい．そのため，残響が存在する環境と同様に音声の特徴が歪みやすく，音声の明瞭性や了解性の低下につながると言える．以上のことから，実環境下では残響や雑音の影響があるために，音声の不明瞭化が大きくなったと考えられる．本研究では最終的に，時間領域において音声言語情報を直接的に処理した上で，間接的（STI を基準とした方法による残響付与）に処理することで音声プライバシー保護の方法を実現することを想定している．実環

境下では残響や雑音の影響があることから，あらかじめ本手法で目的音声を不明瞭化させた上で残響を付与することで，効果的に音声プライバシーを保護できることが期待できる．

人は両耳で受信した信号を頼りに，音源が位置する方向を知覚できる．そのため，音源がいくつも存在するような環境においても，音源の方向を知覚することで目的音のみを取り出して聴くことができる．したがって，特定の方向に注意を向けて音を聴くことで，その方向に存在する音を聴き取りやすくなる場合がある [39]．さらに，音源が存在する位置を視覚的に知覚することが，音源定位の方向に影響を与えることもわかっている [40]．音の到来方向の違いは，カクテルパーティ効果を働かせる要因の一つである [22]．そのため，分凝して音を知覚する際には，音源の方向情報が重要であるといえる．人は両耳で音を聞くとき，両耳間時間差（ITD: interaural time difference），両耳間レベル差（ILD: interaural level difference）などの両耳情報を利用して音源の方向を定位している．本研究では，被験者の正中面にスピーカを設置し，目的音声と局部時間反転音声の二つの音源が同じ方向から，同じ音圧で聴こえるように調整して，音声刺激を再生した．そのため，音源を定位する際に分凝が働くことなく，二音の知覚的融合が生じ，音声を不明瞭化できたと考えられる．このことから，実環境で音声プライバシー保護をする場合，局部時間反転音声を再生する方向に注意すれば効果的にプライバシーを保護できることが期待できる．

本研究では，目的音声を不明瞭化させるための提示音声として局部時間反転音声を用い，目的音声と同じ音圧レベルで被験者に提示して実験を行った．このとき，目的音声と局部時間反転音声の二つの音を知覚的に融合させ，一つの音として聴取者に知覚させることで音声プライバシーを保護することを狙っている．音情景解析の概念に基づく研究 [11] では，目的とする会話音声を不明瞭化させるための提示音声として，防聴音が利用された．防聴音とは，Bregman の発見的規則 [20] に基づく特徴を持つ音声であり，スペクトル包絡を変形することで作成される．この研究では，会話音声を 50 dB，防聴音を 44, 50, 56 dB で提示して単語了解度試験が行われた．その結果，会話音声と防聴音を同じ 50 dB の音圧レベルで提示した場合の平均単語正答率は約 60%であった．さらに，防聴音を 56 dB で提示した場合には平均単語正答率が約 20%まで低下した．このことから，防聴音の提示音圧を大きくすることで音声プライバシーを保護できることがわかった．しかし，過度なスペクトル包絡の変形により作成された防聴音は不快感を生じさせることがあるため，大きい音圧で提示することは望ましくないと考えられる．本研究では実環境において，低親密度の場合には最大で 20%まで，高親密度の場合には最大で 48%まで単語正答率が低下した．このことから，本手法を用いた場合には，提示音声の音圧を大きくせずに目的音声を不明瞭化できると言える．したがって，不快感やマスキング法で問題となっている騒音暴露などを生じさせず，効果的に音声プライバシーを保護できることが期待できる．

第7章　結論

7.1　本研究で明らかにしたこと

本研究では，時間領域において音声言語情報を直接的に処理し，効果的な音声プライバシー保護の方法を実現することを目的とした．そこで，振幅包絡線包絡線情報のみを局部時間反転した音声を目的音声に知覚的に融合させ，目的音声を不明瞭化できるかどうかを検討した．その結果，明らかになったことを以下に示す．

- 対象となる目的音声の時間微細構造はそのままで，振幅包絡線情報のみを局部時間反転した音声を目的音声に知覚的に融合させることで，目的音声を不明瞭化できる．
- 局部時間反転の反転区間長を長くすることで，目的音声の不明瞭化が大きくなる．
- 反転区間長を長くすると，不明瞭化の度合いがおおよそ一定となる点がある．
- 実環境下でも，反転区間長の操作によって音声を不明瞭化できる．

これらの結果から，音声の振幅包絡線情報を局部時間反転することで，目的とする音声を不明瞭化できることが明らかになった．また，局部時間反転によって音声言語情報を壊すことができ，局部時間反転処理が音声の時間構造の操作として有効な処理であることが明らかになった．以上のことから，時間領域において音声言語情報を直接的に処理することで，効果的に音声プライバシーを保護できることが示された．

7.2　残された課題

- STI を基準とした音声プライバシー保護の方法との組み合わせ

  本研究では，時間領域において音声言語情報を直接的に処理した音声を，知覚的融合を促進する形で目的音声に付与することで目的音声を不明瞭化できることが示された．また，実環境下で本手法を用いた場合，残響や雑音の影響によって目的音声の不明瞭化が大きくなることが予想された．これらのことから，あらかじめ本手法で目的音声を不明瞭化させた上で，STI を基準と

した方法 [8] によって残響を付与することで，より効果的に目的音声を不明瞭化できると考えられる．二つの方法を組み合わせることで，STI を低下させすぎることなく音声プライバシーを保護でき，目的音声の聴き取りにくさの担保と，わずらわしさの低減の両立が期待できる．さらに，反転区間長を長くしすぎることなく効果的な音声プライバシー保護の方法を実現できると考えられる．

- 知覚的融合が生じる条件の調査

  本研究では，目的音声とその振幅包絡線情報を局部時間反転した音声を同時に再生して検討を行った．しかし，実際のオープンスペースにおいて本手法を用いる際には，目的音声を収録し反転して再生することとなる．そのため，局部時間反転音声を再生する際の遅延を考慮しなければならない．スペクトル包絡の変形によって防聴音を作成した場合，基本周波数の違いや音の立ち上がりのずれが知覚的融合に影響を及ぼすことが明らかになっている [41]．今後，振幅包絡線情報を局部時間反転した音声を用いた場合には，目的音声との立ち上がりのずれがどの程度までであれば知覚的に融合するかについて検討する必要がある．

参考文献

[1] Denes, P.B., Pinson, E.N., “The Speech Chain: The Physics and Biology of Spoken Language,” 2nd ed., W. H. Freeman, New York, 1993.

[2] 荒井隆行, “音声コミュニケーションにおける Speech Chain を考える,” 情報処理学会研究報告, vol. 115, no. 3, pp. 1–2, 2017.

[3] 前川喜久雄, 北川智利, “音声はパラ言語情報をいかに伝えるか,” 認知科学, vol. 9, no. 1, pp. 44–46, 2002.

[4] Fujisaki, H., “Prosody, modeles, and spontaneous speech, in Computing Prosody,” Y. Sagiska, N. Campbell, and N. Higuchi(Eds), Springer, pp. 27–42, 1996.

[5] 佐藤 洋, 清水 寧, “スピーチプライバシー研究の歴史と近年の動向,” 日本音響学会誌, vol. 64, no. 8, pp. 475–480, 2008.

[6] Cavanaugh, W. J., Farrel, W.R., Hirtle, P. W., and Watters, B. G., “Speech Privacy in Buildings,” J. Acout. Soc. Am., vol. 34, no. 4, pp. 475–492, 1962.

[7] IEC 60268-16:2003. “Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index”.

[8] 柏原 佑太, “音声伝送指標を基準としたスピーチプライバシー保護の研究,” 北陸先端科学技術大学院大学修士論文, 2017

[9] 宮崎 晃和，森田 翔太，鵜木 祐史, “背景雑音を考慮した音声伝送指標のブラインド推定法の検討,” 電子情報通信学会技術研究報告, vol. 113, no. 349, pp. 1–6, 2013.

[10] 鵜木 祐史，佐々木 恭平，宮内 良太，赤木 正人, “残響音声からの音声伝達指標のブラインド推定法の検討,” 電子情報通信学会技術研究報告, vol. 113, no. 134, pp. 63–68, 2013.

[11] 赤木 正人, 入江 佳洋, “音情景解析の概念にもとづいた音声プライバシー保護,” 電子情報通信学会論文誌 A, vol. J97-A, no. 4, pp. 247–255, 2014.

[12] Drullman, R., “ Temporal envelope and fine structure cues for speech intelligibility,” J. Acoust. Soc. Am., vol. 97, no. 1, pp. 585–592, 1995.

[13] Ueda, K., Nakajima, Y., Ellermeier, W. and Kattner, F., “Intelligibility of locally time-reversed speech: A multilingual comparison,” Scientific reports, vol. 7, no. 1, pp.1–8, 2017.

[14] Matsuo, I., Ueda, K. and Nakajima, Y., “Intelligibility of chimeric locally time-reversed speech,” J. Acoust. Soc. Am. Express Letters, vol. 147, EL523-EL528, 2020.

[15] Ueda, K., Nakajima, Y., Kattner, F. and Ellermeier, W., “Irrelevant speech effects with locally time-reversed speech: Native vs non-native language,” J. Acoust. Soc. Am., vol. 145, no. 6, pp. 3686–3694, 2019.

[16] 李孝珍, 上野佳奈子, 坂本慎一, “調剤薬局におけるスピーチプライバシーの改善事例に関する実験的検討,” 日本建築学会技術報告集, vol. 20, no. 44, pp. 165–168, 2014.

[17] 佐伯 徹郎, 藤井 健生, 山口 静馬, 老松 健成, “音声をマスクするための無意味定常雑音の選定,” 電子情報通信学会論文誌 A, ,vol. J86-A, no. 2, pp. 187–191, 2003.

[18] 佐伯 徹郎, 山口 静馬, 為末 隆弘, “マスキングノイズによるスピーチプライバシー保護に関する一考察,” 日本音響学会誌, vol. 61, no. 10, pp. 571–575, 2005.

[19] 星野 康, 森本 政之, 佐藤 逸人, “遮音性能とスピーチプライバシーの関係,” 日本建築学会講演論文集, D–1, pp. 343–344, 2010.

[20] Bregman, A. S., “Auditory scene analysis: hearingin complex environments,” in Thinking in sound:The cognitive psychology of human audition, ed. S.McAdams and E. Bigand, Chapter 2, Oxford SciencePub., pp. 10–36, 1993.

[21] Cherry, E. C., “Some experiments on the recog-nition of speech, with one and with two ears,” J. Acoust. Soc. Am., vol. 25, no. 5, pp. 975–979, 1953.

[22] 赤木正人, “カクテルパーティ効果とそのモデル化,” 電子情報通信学会誌, vol. 78, no. 5, pp. 450–453, 1995.

[23] Bregman, A. S., “Auditory scene analysis: the perceptual organization of sound,” MIT Press, Cambridge, MA, 1990.

[24] Robert V. Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, Michael Ekelid, “Speech Recognition with Primarity Temporal Cues,” Sciense, vol. 270, Issue 5234, pp. 303–304, 1995.

[25] Tachibana, R. O., Sasaki, Y. and Riquimaroux, H., “Relative contributions of spectral and temporal resolutions to the perception of syllables, words, and sentences in noise-vocoded speech,” Acoust. Sci. Tech., vol. 34, no. 4, pp. 263–270, 2013.

[26] Ueda, K., Araki, T. and Nakajima, Y., “Frequency specificity of amplitude envelope patterns in noisevocoded speech,” Hearing research, vol. 367, pp. 169–181,2018.

[27] Loizou, P. C., Dorman, M., and Tu, Z., “On the number of channels needed to understand speech,” J. Acoust. Soc. Am., vol. 106, no. 4, pp. 2097–2103, 1999.

[28] Xu, L., Pfingst, B. E., “Spectral and temporal cues for speech recognition: Implications for auditory prostheses,” Hearing research, vol. 242, pp. 132–140,2008.

[29] Licklider, J. C. R., Miller, G. A., “The perception of speech,” Handbook of Experimental Psychology, pp.1040–1074, 1951.

[30] Meyer - Eppler, W., “Reversed speech and repetition systems as means of phonetic research,” J.Acoust. Soc. Am., vol. 22, no. 6, pp. 804–806, 1950.

[31] 日本音響学会編, 聴覚モデル, コロナ社, pp. 101–115, 2011.

[32] 山本克彦, 入野俊夫, 松井淑恵, 荒木章子, 木下慶介, , 中谷智広, “動的圧縮型ガンマチャープフィルタバンクを用いた音声明瞭度予測法: 強調音声を対象とした比較検討.,” 情報処理学会研究報告, vol. 111, no. 20, pp. 1–6, 2016.

[33] Patterson, R., Nimmo-Smith, L,, Holdsworth, J. and Rice, P., “An auditory filter bank based on the gammatone function,” Paper presented at a meeting of the IOC Speech Group on Auditory Modelling at RSRE, pp. 14–15, 1987.

[34] 近藤 公久, 坂本 修一, 天野 成昭, 鈴木 陽一, “親密度別単語了解度試験用音声データセット 2007(FW07) の作成,” 電子情報通信学会技術研究報告, vol. 107, no. 432, pp. 43–48, 2008.

[35] 坂本 修一, 天野 成昭, 鈴木 陽一, 近藤 公久, “単語了解度試験におけるモーラ同定に対する親密度の影響,” 日本音響学会誌, vol. 60, no. 7, pp.351–357, 2004.

[36] 吉岡拓也, 中谷智広, “確率モデルを用いた音声強調: 雑音抑圧, 音源分離, 残響除去, 統合技術及びその応用 (¡ 小特集¿ 近年の音響信号処理における数理科学の進展),” 日本音響学会誌, vol. 68, no. 11, pp. 572–577, 2012.

[37] 橘秀樹, “室内音響測定の現状と今後の課題,” 日本音響学会誌, vol. 49, no. 2, pp. 97–102, 1993.

[38] 森田 翔太, “音環境バリアフリーのためのパワーエンベロープ処理体系,” 北陸先端科学技術大学院大学博士論文, 2017.

[39] Ebata, M., and sone, T., “Improvement of hearing ability bydirectional information,” J.Acoust. Soc. Am., vol. 43, no. 2, pp. 289–297, 1968.

[40] 西田 鶴代，筧 一彦，穂刈 治英，島田 正治, “音源定位における視覚情報の影響: FLMP による視覚情報の影響の定量化,” 日本音響学会誌, vol. 55, no. 11, pp. 735–741, 1999.

[41] 蓑輪 明子, “音声の知覚的融合が生じる条件に関する基礎的研究,” 北陸先端科学技術大学院大学修士論文, 2007.

謝辞

本研究の遂行にあたり，厳しくも丁寧なご指導とご助言を賜りました指導教官の鵜木祐史教授に深く感謝いたします．また，研究室会議をはじめ，様々な機会でご助言を賜りました赤木正人教授に心より感謝いたします．さらに，研究や実験に関するご助言を賜りました木谷俊介助教，小林まおり博士に感謝いたします．また，お忙しい中実験にご参加いただいた被験者の皆様には心よりお礼申し上げます．そして公私にわたりお世話になりました赤木・鵜木研究室の皆様に改めてお礼申し上げます．最後に，音を志すことを決意した日から 10 年間，信念を貫き通させてくれた家族に心から感謝を申し上げます．

研究業績

国内学会における発表

（口頭，査読無）

1. 坂本 貴望, 鵜木 祐史, “時間反転音声と知覚的融合に関する検討,” 2020 年度電気・情報関係学会北陸支部連合大会, G-1, 2020.
2. 坂本 貴望, 小林まおり, 鵜木 祐史, “振幅包絡線情報の局部時間反転による音声の不明瞭化の検討,” 聴覚研資, 50(6), 321–326, 2020.
3. 坂本 貴望, 小林まおり, 鵜木 祐史, “振幅包絡線情報の局部時間反転による音声プライバシー保護の検討,” 音講論 (春), 3-4P-4, 2021.

その他の業績

（受賞）

1. 坂本 貴望, 学生優秀論文発表賞, 2020 年 9 月.
2. 坂本 貴望, 日本音響学会北陸支部優秀学生賞, 2021 年 3 月.

付 録A　局部時間反転音声の知覚的融合に関する検討

A.1　実験目的

二つ以上の異なる音が，ある条件のときに一つの音として知覚されることを知覚的融合という．Bregmanは，知覚的融合が生じる際の条件として,次の四つの発見的規則:（1）共通の立ち上がり/立ち下がりの規則,（2）漸近的変化に関する規則,（3）調波関係に関する規則,（4）共通運命の原理に関する規則を説明した．一方，聴覚の変調知覚の側面から，音声の言語・非言語知覚に関して，振幅包絡線情報（TAE）と時間微細構造（TFS）の重要性が検討されている．これらの特徴に関し，二つの音が知覚的に融合する条件を明らかにできれば，変調知覚における特徴の時間構造について議論できるかもしれない．そこで，局部時間反転音声を用いた知覚的融合の実験を行うことで，この疑問点の解明に取り組む.

A.2　実験方法

実験では，発見的規則（1）,（2）,（4）のいずれを満たした場合に原音声とその局部時間反転音声の二つの音が知覚的に融合するかを調査した．ここでは次の三つの条件:規則（1）を満たす条件として原音の音声区間内（SPS）でのみ局部時間反転した場合，規則（3）を満たす条件としてTAEのみを局部時間反転した場合，規則（4）を満たす条件としてTFSのみを局部時間反転した場合を検討した．また，リファレンスとして原音の全区間（ALL）で局部時間反転した場合，ならびにTAE・TFSの両方を局部時間反転した場合についても検討した.

A.2.1　被験者

実験には，日本語を母語とし正常聴力を有する成人10名（22-29歳，男性8名，女性2名）が参加した.

A.2.2　装置と刺激

実験刺激は，PC（LG Sharkoon, Windows10）より，A/D コンバータ（Steinberg UR44），およびヘッドホンアンプ（STAX SRM-1）を経由して開放型ヘッドホン（STAX SR-L700）から被験者に提示した．被験者の反応の取得には MATLAB にて作成した GUI アプリケーションを使用し，入力装置にはマウスを使用した．

音声刺激として，男女各 5 名の話者の発話音声と，これらを原音として作成した局部時間反転音声を提示した．時局部時間反転音声は，TAE，TFS，TAE・TFS を時間反転させる 3 条件と，SPS と ALL の 2 条件をそれぞれ組み合わせたものを使用した．局部時間反転の反転区間長は 5，10，20, 40, 80, 160, 320, 640 ms の 8 条件とした．刺激の総数は，480 個（=3 × 2 × 8 × 10 音声）であった．音声刺激は，人工耳（BK Artificial Ear Type 4153），マイク（BK Microphone Type 4192），騒音計（BK Sound Level Meter Type 2250）を用いて，A 特性音圧レベルがおよそ 62 dB となるよう設定した．

A.2.3　手続き

単語了解度試験は防音室で行われた．被験者には，目的音声である発話音声に局部時間反転音声を加算して，ランダムな順序で提示した．被験者の課題は，聴き取った音声が知覚的に融合したかどうかを，二肢強制選択することだった．実験に要した時間は休憩を含め 1 時間程度であった．

A.3　実験結果

図 A.1 に，TAE，TFS，TAE・TFS ならびに，SPS，ALL の条件ごとの二音の知覚的融合の結果を示す．TAE，TFS，TAE・TFS の 3 条件では，TAE の融合率が最も高く，反転区間長 5〜40 ms で融合率が 90%となることがわかった．また，SPS と ALL の条件では，全般に音声区間に限定して時間反転するほうが知覚的融合に効果がみられるようである．以上から，原音声とその局部時間反転音声の知覚的融合に関しては，発見的規則（調波性に関する規則）が最も重要であることがわかった．

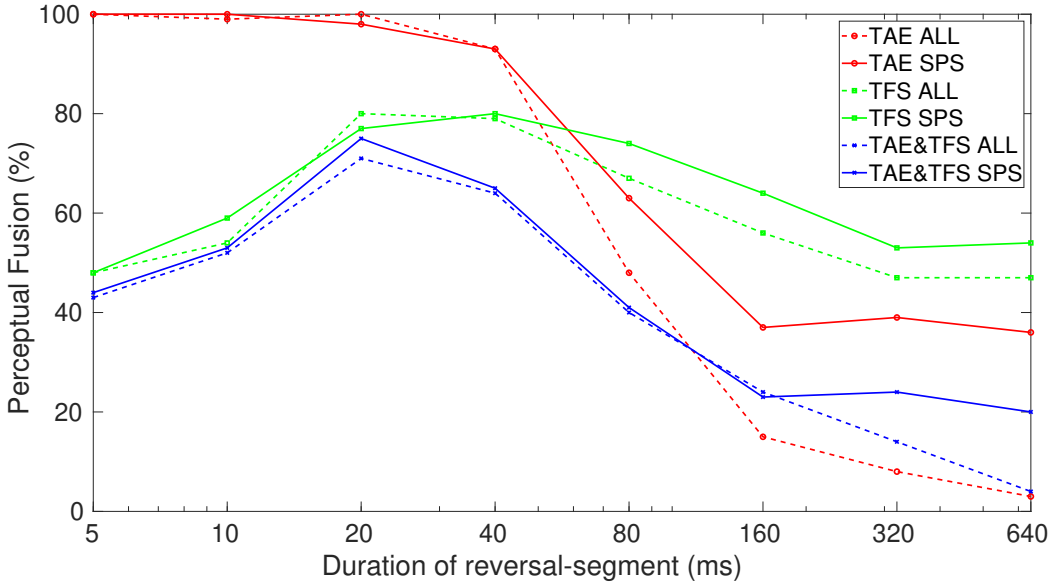

図 A.1: 局部時間反転音声の融合率