修士論文

変調スペクトルに着目した残響環境下での音声了解度向上に関する研究

1810183　森田 翔太

主指導教員　赤木 正人
審査委員主査　赤木 正人
審査委員　鵜木 祐史
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 02 年 03 月

Abstract

Reverberation various reverberation times exists in indoor public spaces such as stations and concert halls. In the presence of reverberation such as evacuation guidance voices during disasters, it is necessary to accurately transmit voices to people in the facility.

When speaking in noisy environments, humans change voice in response to noise to improve speech intelligibility. This phenomenon is known as the Lombard effect. Similar compensation movements may be performed in reverberant environments. To date, various studies have been conducted on speech in reverberant environments. However, it has not been clarified yet features that are important for accurately transmitting speech in the deformation of speech in a reverberant environment.

Arai et al. that speech uttered in a reverberant environment tended to have higher intelligibility in a reverberant environment than speech uttered in a silent environment. In addition, they report that consonants are coordinated by steady-state suppression corresponding to a vowel for speech, and the intelligibility under reverberation is significantly increased.

Kubo et al. speech uttered under a total of four conditions of reverberation environment of quiet sound and reverberation time (1 second, 3 seconds, 5 seconds), and conducted a listening experiment with reverberation convolved. It was reported that some speakers tended to have higher speech intelligibility in reverberant environments for longer. Also, focusing on the formant frequency of the uttered voice of the same speaker, analysis is performed, and the vowel space is expanded when the reverberation time during utterance becomes longer, and the first formant in the onset (0 to 25 %) of the vowel section The slope of the transition between (F1) and the second formant (F2) is the steepest in the non-giving condition, but the transition of the formant is steeper when the reverberation time is 5 seconds compared to the reverberation time of 1 second when speaking. Has been reported.

However, in each case, only the results on the frequency axis are taken into account, and the time variation of the utterance is not considered. In this paper, temporal envelope of speech in order to clarify acoustic features that are important for improving speech intelligibility in reverberant environments. By analyzing the modulation spectrum of speech in various reverberation environments, focusing on F1 and F2, which are considered important in previous research,This paper was found features that are considered to be related to the improvement of speech intelligibility. A listening experiment is performed to determine whether the obtained features have contributed to the improvement of speech intelligibility in a reverberant environment.

Speech uttered in a space resembling a reverberation environment with the same reverberation time (T60) as Schroeder's RIR to investigate whether humans are deforming speech to compensate for information lost by reverberation Speech (40 voices per condition), whose intelligibility tended to increase in the reverberant environment corresponding to the extension of the reverberation time (1 s, 3 s, 5 s), was used for analysis. The frequency bands corresponding to vowels F1 and F2 were extracted with a bandpass filter, and the modulation spectrum was derived to obtain the average value of 40 samples. The results showed that the modulation spectrum of a speech with a long reverberation time tends to be larger than that of a speech with a short reverberation time. In addition, the modulation spectrum around the modulation frequency of about 10 Hz is particularly large, and it is considered that there is a possibility that compensation motion is performed for the modulation frequency component that is damaged when the reverberation time is long.

Next, in order to evaluate the relationship between the utterance deformation and the modulation spectrum in the reverberant environment obtained by the analysis, the intelligibility survey was conducted by listening experiments. The stimulus sound creation method used in the listening experiment was resampled to a sampling frequency of 16000 Hz, and then divided into 64 channels (Band-Width: 125 Hz) using an acoustic filter bank. After obtaining the power envelope from the output of each channel and filtering it to 0 and 1 using Voice Activity Detection (VAD), it is further divided into 64 channels (BandWidth: 1.95Hz) by the Modulation filterbank and specific modulation by the gain control The frequency component is raised and returned to voice. This time, five native speakers of Japanese speak the three most important words in the ATR database and use the voices of the three-mora words. The frequency bands corresponding to vowels F1 and F2 (male: 300 to 2000 Hz) A sound (4 dB up) in which the modulation frequency of the power envelope of the filter envelope: 2 to 16 ch) was around 4 and 10 Hz (Modulation filter bank: 2, 3 ch, 5 and 6 ch) was raised by 4 dB (8 dB up) and a sound was raised by 8 dB. For the experimental stimulus, three types of reverberation (1 s, 3 s, 5 s) convoluted with the original voice, 4 dB up, and 8 dB up. An experiment was performed using this speech to determine the intelligibility of words in mora units. Nine native speakers of Japanese in their 20 s participated in the listening experiment. The stimulus sound was blocked at each reverberation time, and the sound in the block was presented randomly.

Although there was no difference in intelligibility between the reverberation time of 5 s and 3 s, the results showed that the intelligibility of 4 dB up was higher than that of the original sound under the condition of the reverberation time of 3 s. From this fact, it was suggested that the utterance deformation of raising

the modulation frequency component for the reverberant environment might be significant, and would lead to the improvement of intelligibility.

This study shows that humans can improve intelligibility by increasing the modulated frequency content of speech in reverberant environments, and perform similar processing to increase speech intelligibility in reverberant environments.

目 次

図 目 次

第1章　序論

1.1　はじめに

日常生活を送る上でヒトは言葉を発することにより，コミュニケーションを行っている．音声コミュニケーションにおいて言語情報は重要でコミュニケーションが行われる音環境は様々に変化するため，言語情報が損なわれ音声了解度が低下する場合がある．そのため，ヒトは音環境の変動に対して音声了解度を確保するために音声生成と知覚による制御を行っていると考えられる．例えば人込みや車が多く通る道など雑音が多い環境で会話するとき，ヒトは会話を明瞭にするために周囲の環境に応じて声を変化させている．これは雑音により，自身が発した音声が聞こえ辛くなるため，発話音声の強度を上げ，発話速度を落とし，基本周波数 (F 0) を高くすることでスペクトルの傾きが大きくなるよう発話を変形させているからである [1]．この現象はロンバード効果と知られている [2]．また，雑音のレベルに対して発話が変化することも観察されており，この発話変形と雑音の関係を議論することで雑音環境下での了解度向上の試みが行われている [3]．

雑音と同様に残響によっても音声了解度は低下する．室内における残響は初期反射音と後部残響音の 2 つの要素から成り立っている (図 1.1 参照)．音声了解度低下の主な原因として後部残響音が先行する音素をマスキングする overlap-masking[4] が挙げられる．後部残響音が直接音に対して 60 dB 減衰するまでの時間を残響時間と呼び，残響時間が長いほど，音声了解度は大きく低下する．

駅やコンサートホールなどをはじめとした室内公共空間には様々な残響時間の残響が存在しているため，音声案内において正確に音声を伝達することが困難な場合がある．そのため，残響環境下においても音声了解度を確保する必要がある．また，残響環境下においてもロンバード効果と似た発話変形が行われており，静音環境下で発話された音声に比べて残響環境下で発話された音声の方が残響環境下における了解度が上がったという報告がある [5]．

ヒトは残響による音声了解度低下を小さくするためにどのような発話変形を行っているのかを確かめる．そこから残響環境で正確に情報を伝達するための手掛かりを探す．

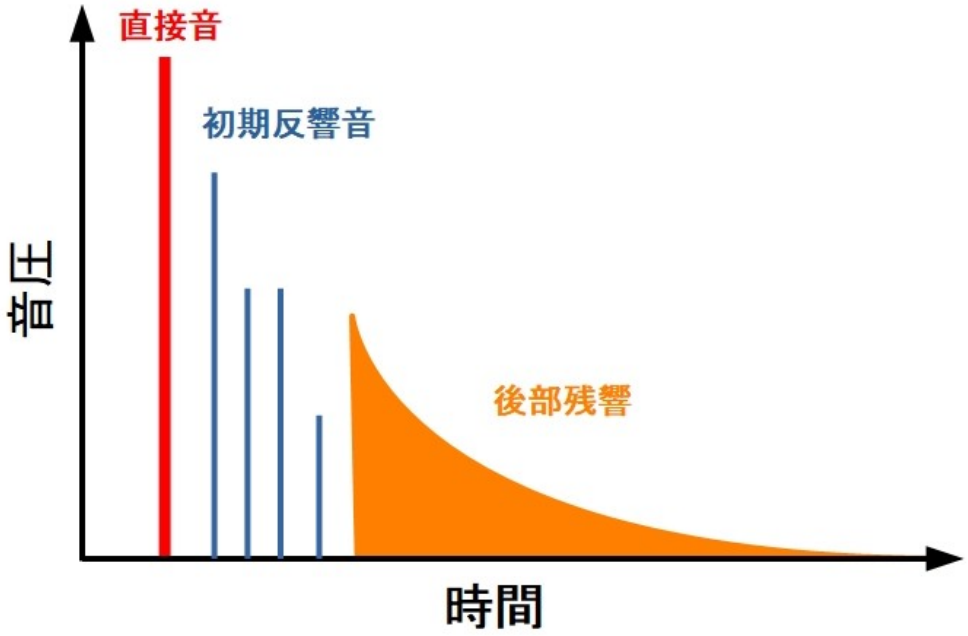

図 1.1: 残響の概略図

1.2 関連研究

残響環境下での発話の了解度について程島らは残響環境下において発話された音声は非残響環境下において発話された音声よりも了解度の低下が小さいことを報告している [5]．また，発話訓練・発話スタイル・話速が残響環境下で聴取者に与える影響について調査し，話速以外は音声明瞭度に関係していることを示した [6, 7]．これにより clear speech は残響環境下で音声明瞭度を上げるために有用であることが分かった．また，雑音環境下で話速を遅くすることが有意であるのに対し，残響環境下では効果が表れない原因は overlap-masking の量も同時に増えることが原因と考えられる．

荒井らは残響による overlap-masking の影響を軽減するため，定常部抑圧処理を行い音声明瞭度改善に繋がったことを報告した [8, 9]．この報告から母音の抑圧処理は overlap-masking による影響を低減させることが分かった．しかし，辻らの研究では母音の抑圧処理と子音強調を行ったところ了解度に有意な差は得られず，母音の抑圧が大きいときに了解度が下がる聴取者がいることを報告している [10]．明瞭度改善に有意であるものの残響環境下で了解度を確保するためには母音の抑圧処理だけでは十分ではないと考えられる．久保らは残響時間が異なる残響環境下で発話された音声について発話時の残響時間が短いときに比べ，長いときの方が残響下での了解度低下が小さくなるという結果を報告している [11]．また，同じ話者の音声において発話時の残響時間が長いとき，母音空間が拡大し，F2 のフォルマント遷移が急峻になるという結果を報告している [12]．これらは conversational と clear speech の差異として知られている特徴である [13]．残響下発話においてフォルマント遷移が急峻であることから発話の時間的包絡線にも影響が出ている可能性があるが，詳しく調査はされていない．

1.3 研究の目的

残響環境下で音声了解度を上げるために重要な音響特徴は明らかになっているとは言えない．また，残響が音声信号の時間構造を損なうにも関わらず，残響環境下での発話変形において音の時間変動を表す変調周波数成分をほとんど調査されていない．

本研究では残響環境下で発話された音声がどのように変動することによって了解度の低下を小さくしているかを調査する．そのために音声の時間的包絡線に着目した上で先行研究 [11, 12] で重要とされている母音の F1 と F2 を中心に様々な残響時間の残響環境下での発話を分析する．その結果から残響環境下での音声了解度向上に関係があると考えられる特徴を見つける．有意と考えられる音響特徴量を変化させた音声を作成し，了解度調査を行うことにより得られた結果が残響環境下において音声了解度を確保するために重要である音響特徴量であるかを確認する．

1.4　本論文の構成

本論文は，5 章で構成される．図 1.2 に概略図を示す．第 1 章は序論であり，本論文での研究の背景と目的，関連研究を述べている．第 2 章は着眼点を述べる．残響が音声了解度に与える影響から残響環境下で起こりうる発話変形を予測している．第 3 章は手法と結果とそれに対する考察を述べており，今回使用した音声についてもこの章で述べる．第 4 章は第 3 章での結果，考察を元に行った了解度調査について述べる．第 5 章は結論であり，本研究により明らかになったことと残された課題について述べる．

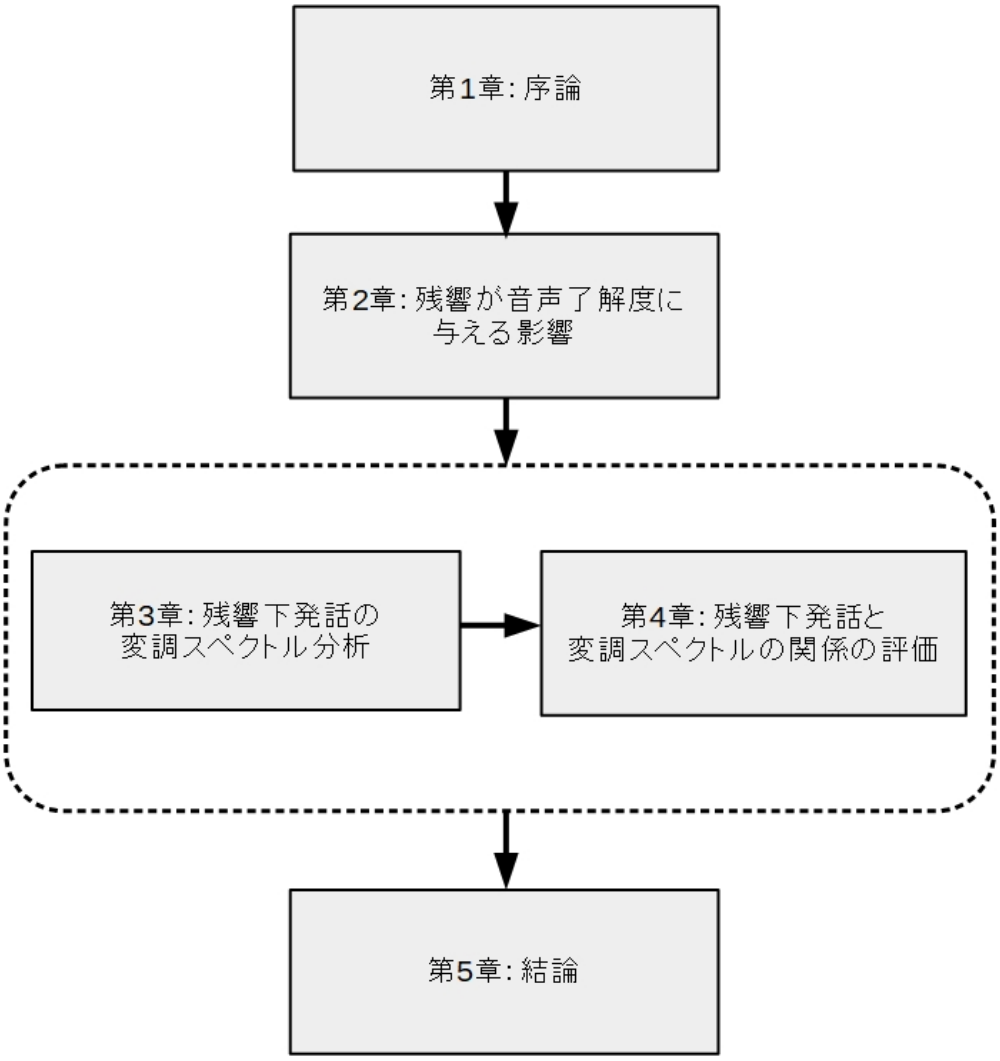
第1章: 序論
第2章: 残響が音声了解度に
与える影響
第3章: 残響下発話の
変調スペクトル分析
第4章: 残響下発話と
変調スペクトルの関係の評価
第5章: 結論

図 1.2: 本論文の構成

第2章　残響が音声了解度に与える影響

2.1　はじめに

残響によって生じる音声了解度低下の原因として後部残響が先行する音素をマスキングしていることが考えられる．後部残響音が直接音に対して 60 dB 減衰するまでの時間を残響時間と呼び，残響時間が長ければ長いほど了解度は低下する．また後部残響音はスペクトル歪みだけでなく波形の包絡線にも影響を与える．

この章では残響の室内インパルス応答 (Room Impulse Response:RIR) モデルから異なる残響時間の後部残響モデルを定義し，その変調伝達関数 (Modulation Transfer Function : MTF) を導出する．残響を模した MTF から音声了解度を低下させている原因を予測することでヒトが残響環境下での発話で行っている可能性がある補償運動について検討していく．

2.2　残響を模した変調伝達関数が発話に与える影響

残響を模した RIR で代表的なものとして Schroeder の RIR があげられる [14]. Schroeder の RIR は次式により定義される．

$$h(t) = aexp(\frac{-6.9t}{T_R})c_h(t) \tag{2.1}$$

Schroeder の RIR モデルを元に作られた人工インパルス応答の波形を図 2.1 に示す．T_R は残響時間を表している．ここで用いた T_R は室内音響指標の一つである T_{60} と同じである．$c_h(t)$ は白色雑音キャリアで a は Power を表している．この RIR と対応する MTF は次式のように表される．

$$m(f_m) = \frac{1}{\sqrt{1 + (2\pi f_m \frac{T_R}{13.8})}} \tag{2.2}$$

f_m は変調周波数である．この式により本研究で用いた 1 秒，3 秒，5 秒の残響時間を持つ残響の MTF を図 2.1 に示す．図より残響を模した MTF は変調周波数の高域を減衰させる特性を示しており，残響時間が長い時ほど変調周波数成分は大きく減衰していることがわかる．これにより音声了解度に重要な情報を持つと考え

られている 16 Hz 以下の変調周波数の成分が大きく損なわれている．ヒトはこの変調周波数成分を持ち上げるような発話変形を行い，残響下での音声了解度を上げている可能性がある．

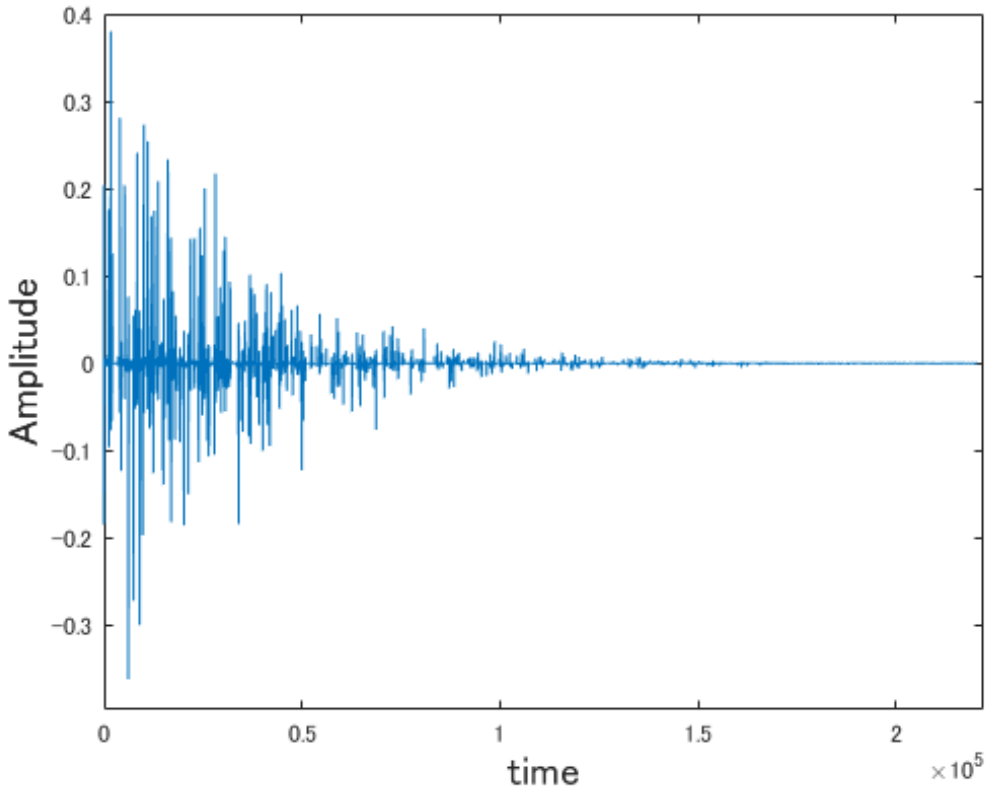

図 2.1: Schroeder の RIR を元に作られたインパルス応答の波形 ($T_R = 5s$)

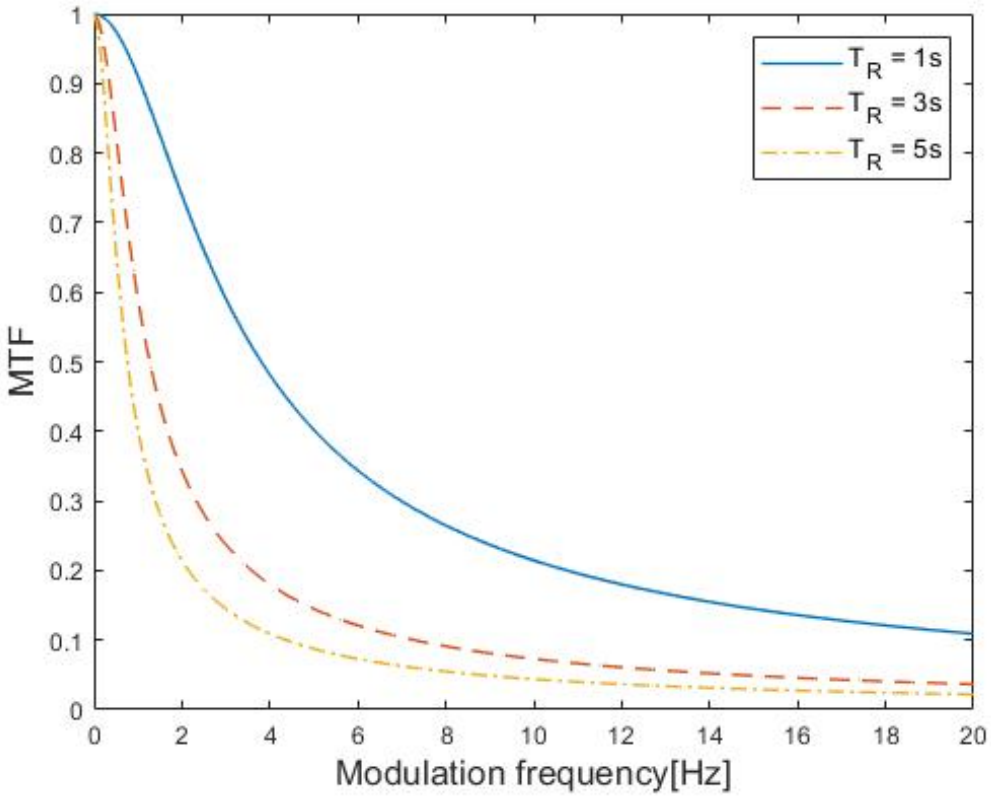

図 2.2: Schroeder の RIR から算出した MTF の概形

2.3 残響による影響の低減

残響を模したMTFは変調周波数の高域を減衰させる特性があり，残響時間が長い環境であるほど減衰は大きい．この特性によって生じる音声了解度低下を軽減させるためにヒトは発話を変化させることで変調周波数成分の高域を持ち上げている可能性がある．このことを確かめるために異なる残響時間の残響環境で発話された音声と静音環境で発話された音声の変調周波数成分を分析する．分析結果から有意と考えられる差を見つけ，その特徴を用いて音声を合成し，了解度調査を通して残響環境での発話変形と変調スペクトルの関係の調査を行う．

残響環境下で了解度を上げるために有意である変調周波数成分を特定し，引き上げることで残響による影響を低減する．

第3章　残響下発話の変調スペクトル分析

3.1　はじめに

変調スペクトルは，音の時間的包絡線のスペクトル情報である．変調スペクトルにおいて，変調周波数が高い成分はパワー包絡線の速い変動を表しており，変調周波数の低い成分はパワー包絡線がゆっくり変動することを表す．また，音声の変調スペクトルは変調周波数の 4∼5Hz 付近にピークがある [15]．このことからヒトは 1 モーラあたり 200∼250 ms の話速で発話していることが示唆されている [16]．変調スペクトルと音声知覚の関係について Dudley は振幅包絡線の 25 Hz 以上の変調周波数成分を除去しても，音声了解度が低下しないことを示し，低い変調周波数成分が音声の知覚にとって重要であることを示唆した [21]．この重要な変調周波数成分は残響により大きく損なわれている．また小林らは残響によって下がる変調スペクトルの変調周波数成分を持ち上げて元の音声の値に近づけることで音声了解度改善ができることを示唆した [22]．

この章では異なる残響時間の残響環境下で発話された音声の変調スペクトルを分析する．その結果からヒトは残響によって壊された変調周波数成分を回復させるような補償運動を行っているのかを考察していく．

3.2　変調スペクトルの分析

2.2 で述べた残響環境下で変調周波数成分の高域が減衰する特性に対してヒトは発話を変形させることにより，変調周波数成分を引き上げている可能性がある．そこで Schroeder の RIR と同じ残響時間 (T_{60}) の残響環境で発話された音声と静音下で発話された音声を用いて変調スペクトルの分析を行い，発話時の環境に対する変調周波数成分の振る舞いを確認する．結果からヒトは残響に歪んだ情報に対してどのような補償運動を行っているのかを明らかにする．

今回，先行研究 [11, 12] での母音空間の拡大や母音の F 1，F 2 の遷移など周波数軸での変化に対して発話の時間変動はどのように変化しているのか比較，検討を行うため，母音の F 1，F 2 に着目した上で変調スペクトルの分析を行った．

3.2.1 導出方法

図 3.1 にバンドパスフィルタを使用した変調スペクトル分析のブロックダイアグラムを示す．はじめに入力信号である $s(t)$(残響下発話音声) から特定の周波数帯域 (母音の F 1，F 2) の情報を取り出すためバンドパスフィルタを使用する．次にバンドパスフィルタから出力された $x(t)$ を用いて次式からエンベロープ $e(t)$ を求める．

$$e(t) = (LPF(|x(t) + j \cdot Hilbert(x(t))|)) \tag{3.1}$$

エンベロープ $e(t)$ を二乗することでパワーエンベロープ $e^2(t)$ を求め、次式で定義した変調スペクトル $E(f_m)$ を求める．

$$E(f_m) = |DFT(e^2(t))| \tag{3.2}$$

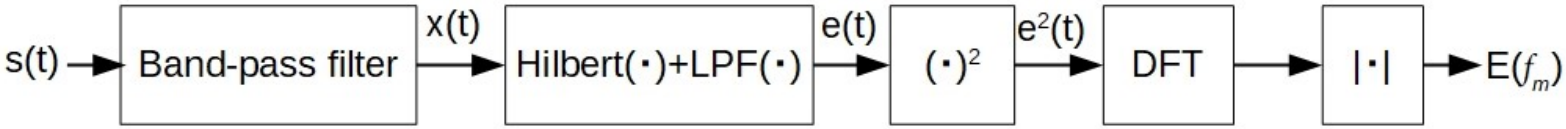

図 3.1: 変調スペクトルの導出方法

3.2.2　音声の収録

今回，先行研究 [11] で発話時の残響時間が長くなるにつれて了解度の低下が小さかったと報告されている日本語母国語話者のデータを使用した．久保らが行った音声の収録方法を図 3.2 に示す．発話時の残響環境の模擬は RIR を用いて作成した残響を話者の発した声に対して，Steinberg Nuendo 7 を用いて畳み込みを行い，開放型ヘッドホン (STAX SR-404) から残響を付与した音声を提示することで実現した．発話環境は 3 つの残響付与条件 (残響時間 (T60) 1 s，3 s，5 s) と 1 つの非付与条件 (non) を模擬した合計 4 条件である．また，音声収録には，雑音混入を極力減らすため超指向性ガンマイク (SENNHEISER 416p 48u) が用いられた．

発話文には FW 07[15] から低親密度である親密度 1.0∼2.5 をターゲット語を挿入した「ここには＿＿と書いてある」という文章を使用した．また，発話文の内容が結果に影響を与えないように各環境で同じ発話文を収録した音声を分析に使用した．

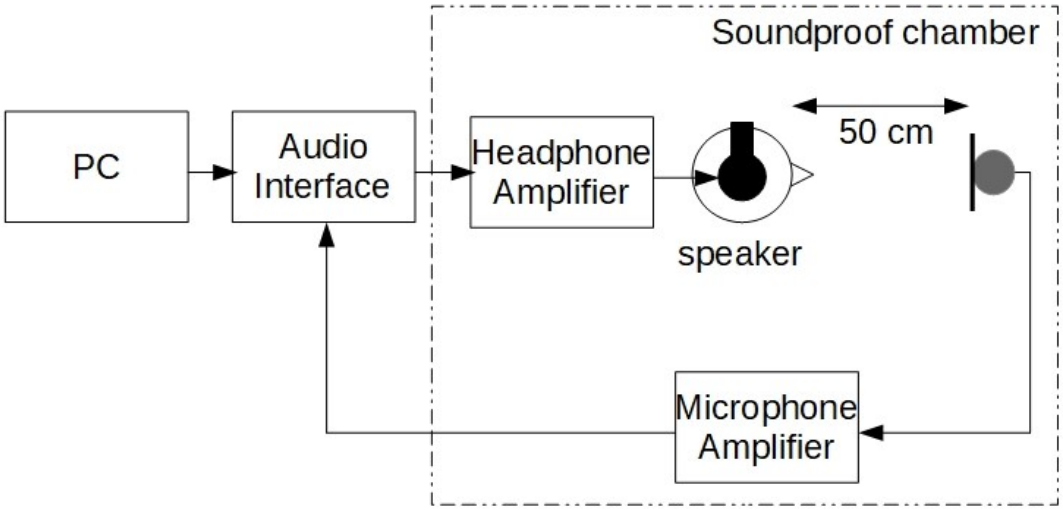

図 3.2: 残響を模擬した空間での音声収録方法

3.2.3　分析の条件

残響による母音空間の拡大 (図 3.3) を考慮した上で男性の母音の F 1，F 2 に対応したバンドパスフィルタ (F 1 : 下限カットオフ周波数 300 Hz，上限カットオフ周波数 550 Hz，F 2:下限カットオフ周波数 800 Hz，下限カットオフ周波数 2000 Hz) を使用した．使用したデータは 3.2.2 で収録された音声 (1 環境につき 40 個) から 4 モーラのターゲット語を取り出した音源であり，サンプリング周波数は 44100 Hz，分析窓幅は 5000 ms で変調スペクトル (40 個の平均) を求めた．

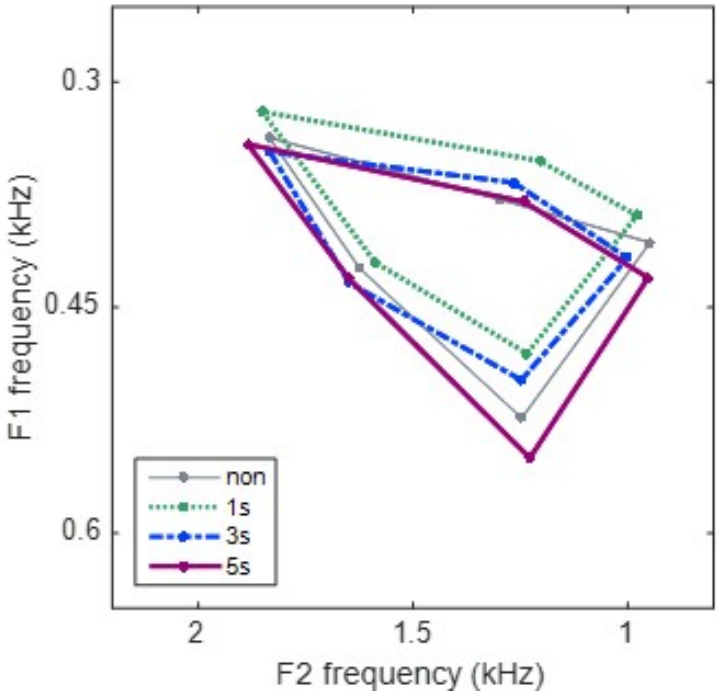

図 3.3: 母音空間の拡大 (出典：[12])

3.3 結果

発話時の残響時間 (non，1s，3s，5s) ごとの変調スペクトル (40 個の平均値) を出力したものを下図 3.4，3.5 示す．図 3.4 は母音の F 1(300∼550 Hz) に着目した変調スペクトルであり，図 3.5 は F 2(800∼2000 Hz) に着目した変調スペクトルである．

母音の F 1 に着目した変調スペクトルにおいて残響時間の長い条件である 3 s と 5 s で発話された音声において変調周波数成分が増幅する傾向が得られた．また，変調周波数の 10 Hz 付近では増幅が顕著に表れている．母音の F 2 に着目した変調スペクトルでは，5 s での発話において変調周波数 10 Hz 付近で他の条件よりも増幅するという結果が得られた．

また，残響時間 1s，3s，5s の条件で発話された女性話者 1 人の発話データを用いて同様の分析を行ったところ，図 3.6，3.7 のような結果が得られた．こちらも残響時間 3s での発話において変調周波数成分が増幅する傾向が得られた．

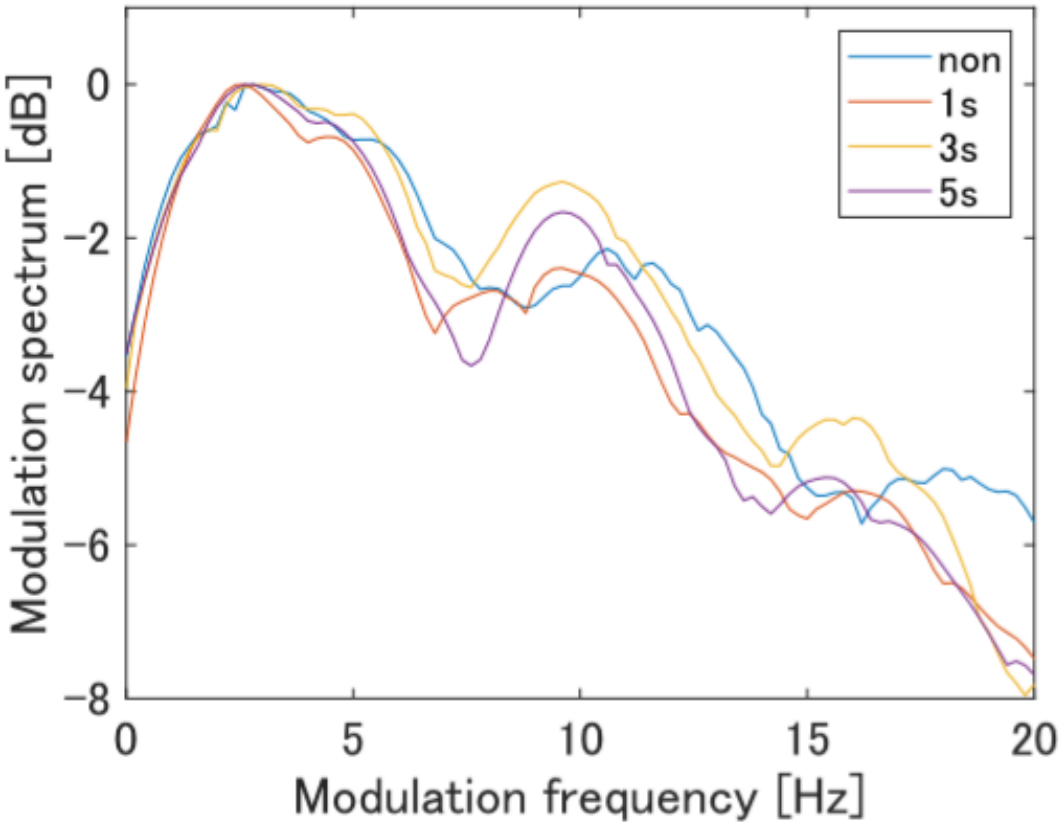

図 3.4: 男性の母音の F 1(300∼550 Hz) に着目した変調スペクトル (凡例は発話時の条件)

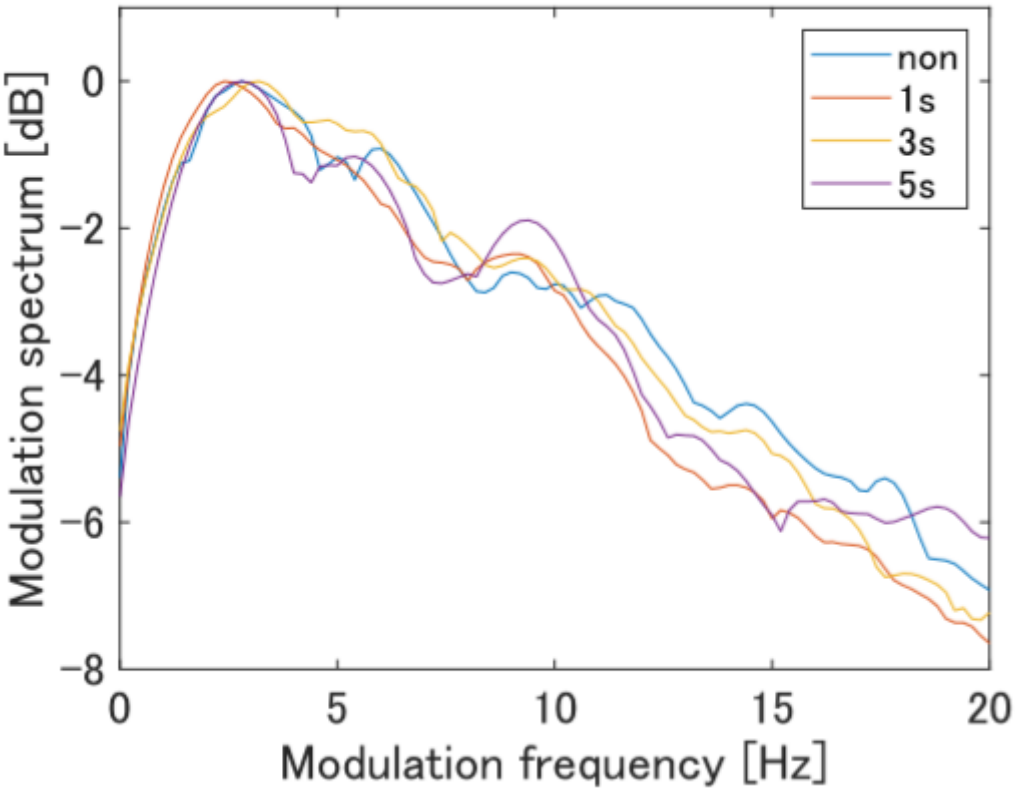

図 3.5: 母音の F 2(800∼2000 Hz) に着目した変調スペクトル (凡例は発話時の条件)

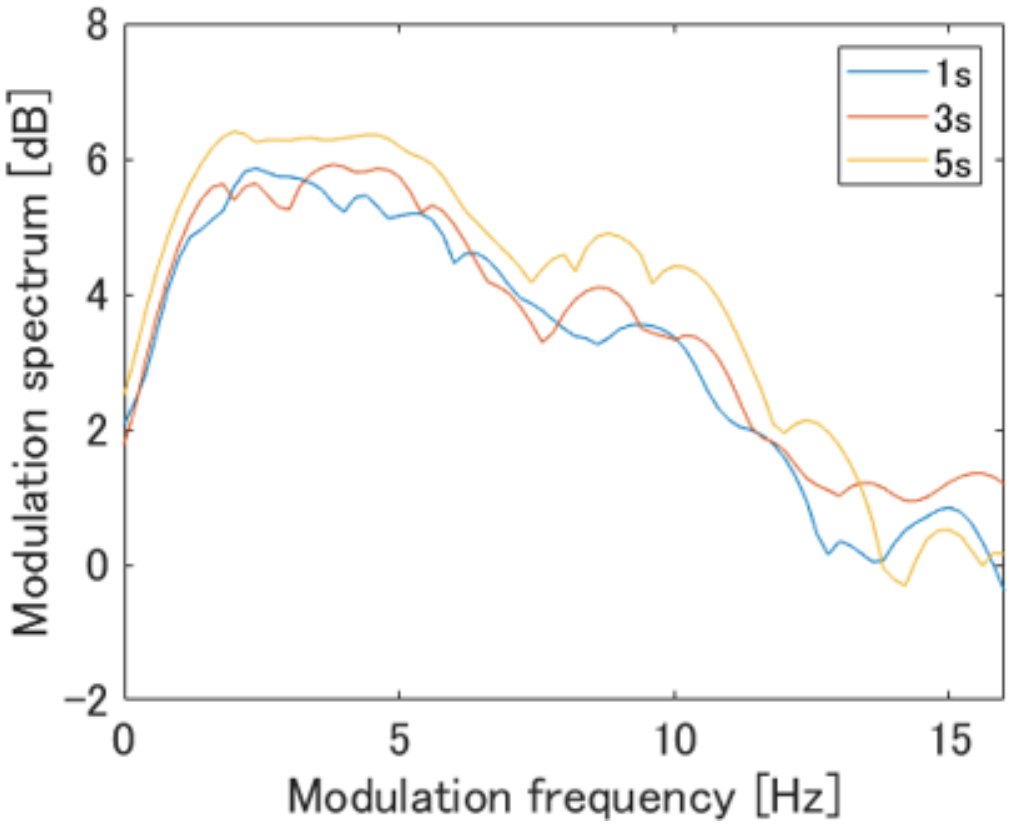

図 3.6: 女性の F 1(250∼1000 Hz) に着目した変調スペクトル (凡例は発話時の条件)

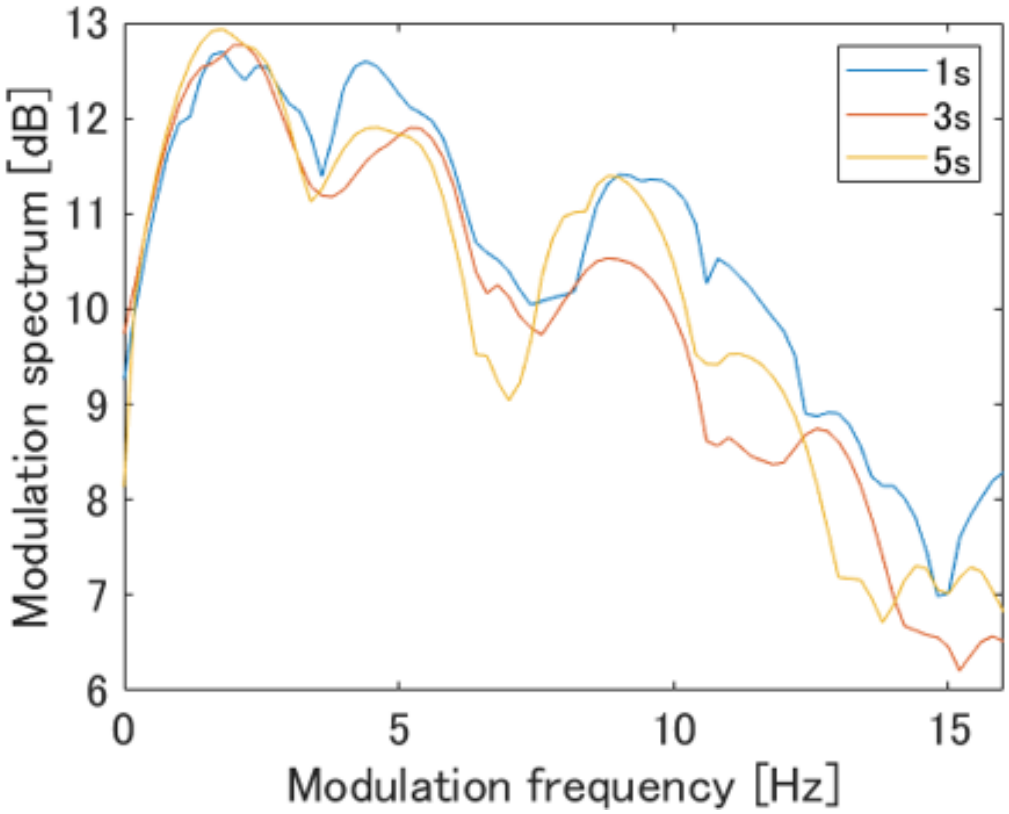

図 3.7: 女性の F 2(800∼3000 Hz) に着目した変調スペクトル (凡例は発話時の条件)

3.4 考察

図 3.4~3.7 より発話時の残響時間が 3 s や 5 s 条件において変調スペクトルの変調周波数成分が全体的に増幅していることがわかる．Drullman らは振幅包絡線において変調周波数成分の調査を行い，変調周波数の 4~16 Hz までの変調成分は音声了解度に重要であると報告している [2][4]．このことから音声了解度において重要とされる変調周波数の 4~16 Hz を持ち上げようとしている可能性が考えられる．

また，図 3.4 で変調周波数 10Hz 付近の変調スペクトルに着目すると発話時の残響時間が 3 と 5 s のときに non や 1 s より大きいピークが得られており，図 3.5 でも 5 s の条件において同様の形状が見られている．話者が女性の場合でも変調周波数 10 Hz 付近で 5s での音声が増幅する傾向が得られている．このことからヒトは残響時間が長い残響環境で変調スペクトルの変調周波数 10 Hz 付近の成分を引き上げるような発話変形をしていることが考えられる．

以上のことから，2.2 で述べた残響を模した MTF によって損なわれる変調周波数成分に対して，ヒトは変調周波数 (特に 10 Hz 付近) の成分を引き上げるような補償運動を行うことで音声了解度を上げている可能性が考えられる．

また，図 3.4，3.5 において母音の F 1，F 2 で結果を比べた時，F 1 に該当する周波数帯域での変調スペクトルの方が残響時間の長さに対する変調周波数成分の増幅が顕著であった．この結果は久保らの先行研究 [12] で報告されている F 2 のフォルマント遷移が急峻であったという結果と逆である．このことから時間的包絡線のスペクトル情報に関しては F 1 付近の周波数帯域に大切な情報がある可能性がある．

第4章　残響下発話と変調スペクトルの関係の評価

4.1　目的

第 3 章で残響時間が長い時の発話において特定の変調周波数が増幅する傾向が得られた．その結果を元に特定の変調周波数成分を持ち上げた音声を作り，了解度調査を行うことで前章で得られた残響環境下での発話変形が音声了解度にどのように関わっているのかを調べる．また，得られた結果から異なる残響時間を持つ残響環境での了解度向上を行う際にどの音響特徴量がどの残響時間において重要であるのかを明らかにしていく．

4.2　実験方法

変調スペクトルにおいて 3.4 で変調周波数 10 Hz 付近の成分を引き上げることが残響時間の長い残響環境で了解度低下を小さくすることに繋がる可能性を示唆した．また，音声の変調スペクトルは変調周波数の 4Hz 付近にピークがあり [15], この変調周波数 4Hz 付近の情報も残響により損なわれている．これらの成分を持ち上げるために変調周波数成分 4 Hz 付近と 10 Hz 付近のパワーエンベロープを引き上げた音声刺激を作り，了解度調査を行っていく．

静音下での発話音声に対して次に述べる手法で変調周波数成分の持ち上げを行い，残響時間 ((T60)1s，3s，5s) の残響を畳み込み日本語を母国語とする 20 代男女に対して聴取実験にて提示した．

4.2.1　実験刺激の作成手法

図 4.1 に実験刺激作成のブロックダイアグラムを示す．原信号 $s(t)$ をアコースティックフィルタバンク [19](図 4.3) を用いて 64 ch の周波数帯域 (帯域幅:125 Hz) のチャンネル信号 $x_k(t)(\mathrm{k} = \mathrm{ch})$ に分解する．次に $x_k(t)$ をヒルベルト変換することでエンベロープ $e_k(t)$ を求める．この時，$x_k(t)$ を $e_k(t)$ で割ることでキャリア $c_k(t)$ を求めておく．$e_k(t)$ を二乗処理することで求めたパワーエンベロープ $p_k(t)$ を変調フィルタバンクに入れる．

図 4.2 に変調フィルタバンクの内部で行っている処理の詳細を示す．まず，発話と関係のない環境音が持ち上げられて振動音のような雑音が発生するのを防ぐために $p_k(t)$ に対して音声区間検出 (VAD) を応用したフィルタリング (O/I フィルタリング : 図 4.4) を行い，重要なパワーエンベロープの成分のみを取り出す．取り出した信号をアコースティックフィルタバンクにより，さらに 64 ch の変調周波数帯域 (帯域幅:1.95 Hz) の $m_k(t)$ に分解し，gain control によって特定の変調周波数を持ち上げる．その後，逆アコースティックフィルタバンクにより変調後のパワーエンベロープ $P_k(t)$ を得る．$P_k(t)$ を $\frac{1}{2}$ 乗することでエンベロープに戻し，求めておいた $c_k(t)$ をかけ合わせることでチャンネル信号に戻す．最後に逆アコースティックフィルタバンクを用いて実験刺激となる $y(t)$ を得る．

今回，母音の F1，F2 に該当する帯域 (300∼2000 Hz) から出力されたパワーエンベロープを用いて変調周波数 4 Hz 付近 (2∼6 Hz) と 10 Hz 付近の (8∼12 Hz) の成分を 4 dB 上げた刺激と 8 dB 上げた刺激を作成した．

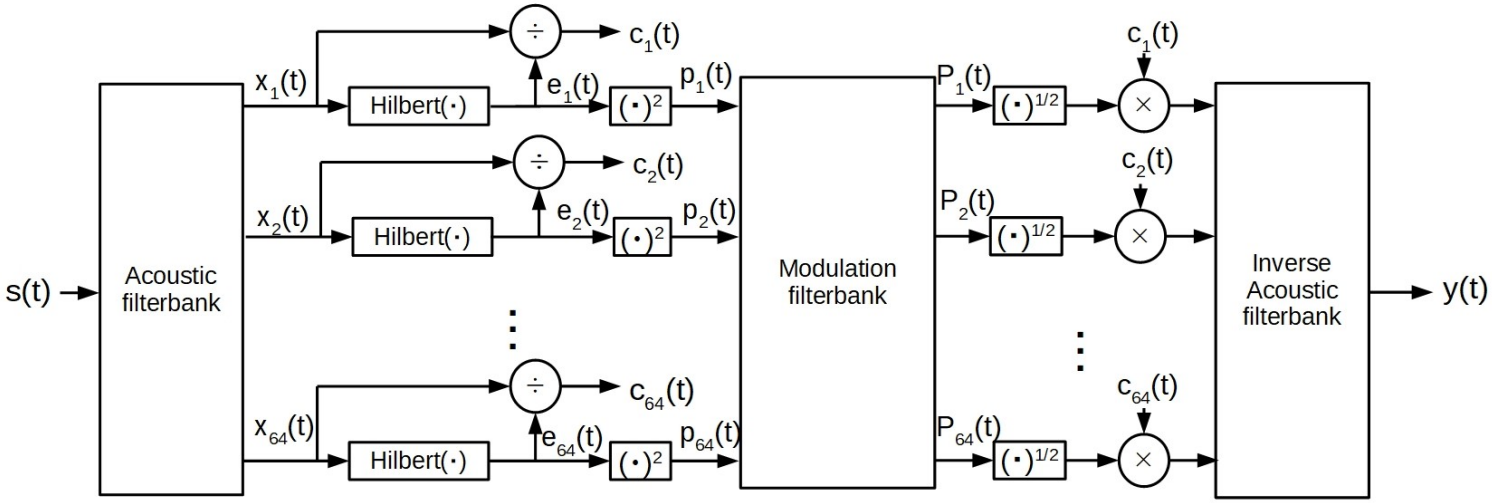

図 4.1: 実験刺激作成のブロックダイアグラム

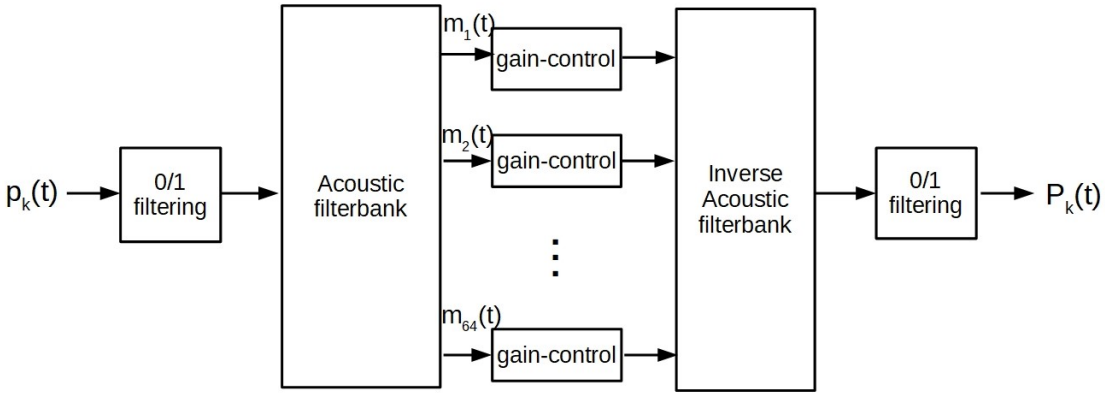

図 4.2: Modulation filterbank 内のブロックダイアグラム (0/1filteringVAD)

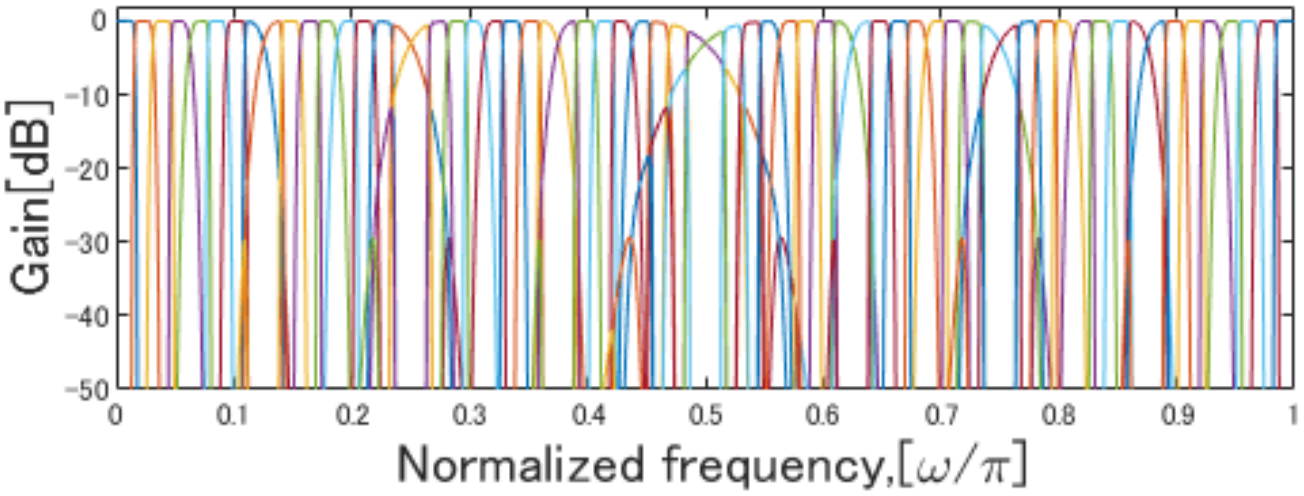

図 4.3: アコースティックフィルタバンクの周波数応答

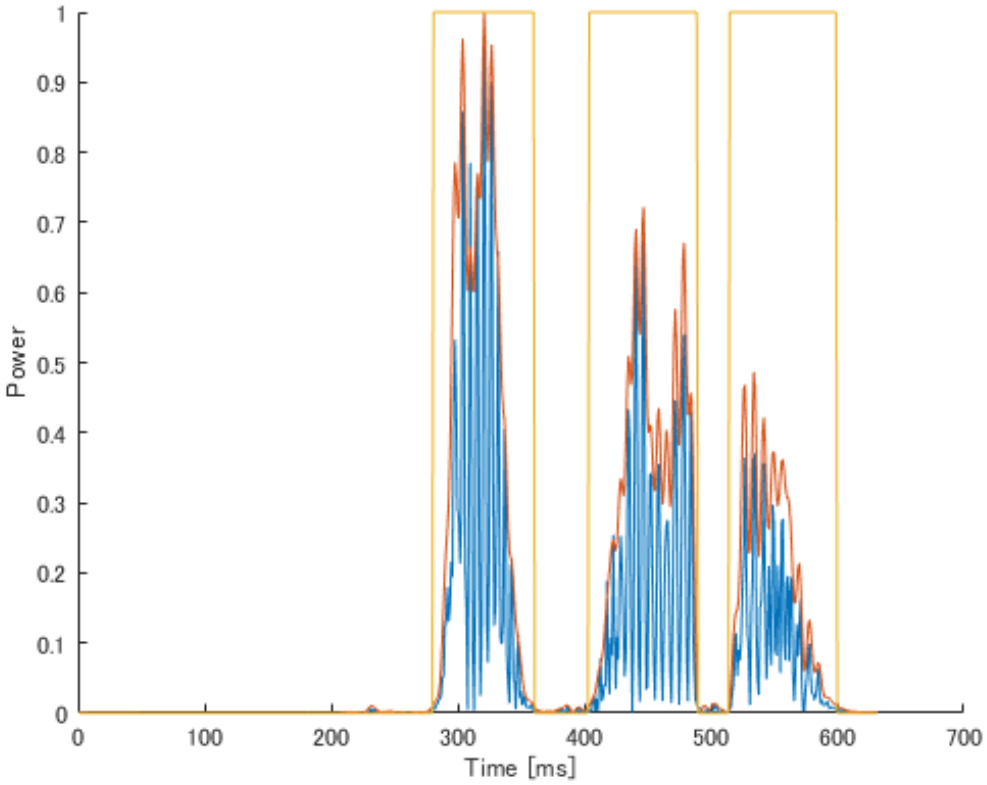

図 4.4: パワーエンベロープに対して行った 0/1 フィルタリングの例

4.2.2　実験の条件

実験には日本語話者 9 人が参加した．実験刺激は日本語母国語話者 5 人が ATR データベース [18] の最重要語 (3 モーラ) を発話したデータから作成した．今回，原音声 (以下：Neutral) と変調周波数 4 Hz 付近と 10 Hz 付近の成分を 4 dB 上げた刺激 (以下：4 dB up) と 8 dB 上げた刺激 (以下：8 dB up) を作った．3 種類の刺激を挿入したコーパス文「アイダ　○○○　アオイ」に対し，3 つの残響時間 ((T60)1 s，3 s，5 s) の残響を畳み込んだ音声 (合計 9 条件 (1 条件当たり 20 個)) を聴取者に提示し，図 4.5 の GUI を用いて聞き取り調査を行った．このとき，発話データによる了解度のかたよりを防ぐために各条件で同じ単語は使用せず，複数の聴取者に同じ単語が同条件で提示されないようにした．また，実験刺激は残響時間ごとにブロック化し，聴取者に対してブロック内の音声をランダムに提示した．

使用した機材は PC：LG sharkoon，ヘッドフォン：STAX SR-L500，オーディオインターフェース:Fireface UCX，モニター:I-O DATA LCD-MF244 であり，音声の提示には MATLAB を用いた．

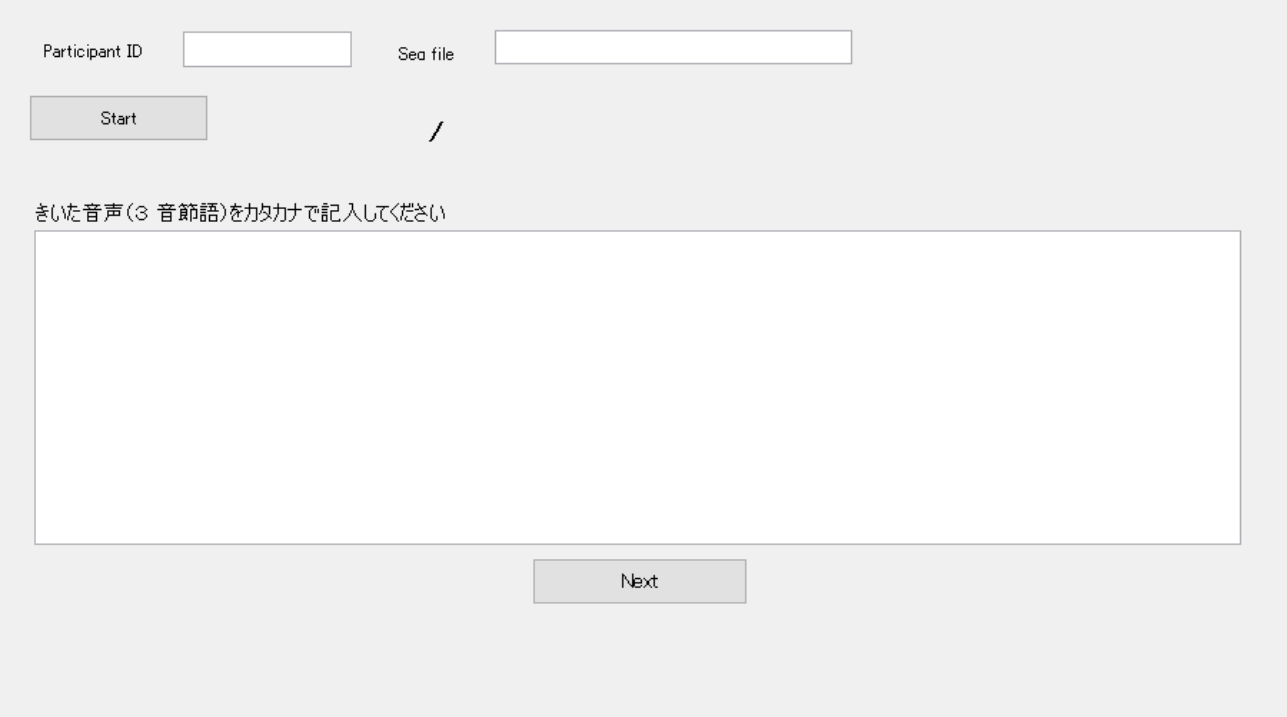

図 4.5: 実験に用いた GUI

4.3 結果

図 4.6 に各条件のモーラ別正答率の平均値を示す．残響時間 1 s では原音，4 dB up，8 dB up の 3 条件とも了解度が高く，差はみられなかった．残響時間 3 s では 4 dB up の正答率の平均値が原音より 6 % 高くなったが，8 dB up では差が見られなかった．条件残響時間 5 s ではどの条件においても了解度が低く明確な差が見られなかった．残響時間 3 s の Neutral と 4 dB up において t-検定を行ったところ，有意水準 5% で有意という結果が得られた．(P 値:0.048)．

4.4 考察

図 4.6 より残響時間 3 s の残響環境において変調周波数成分 4 dB を上げた音声の了解度が，残響環境における了解度向上に関係していることが示唆された．また，8 dB 上げた音声に明確な了解度向上の傾向が見られない点については第 3 章の変調スペクトルの分析結果より残響環境での発話変形においてヒトは 2∼4 dB までしか変調周波数成分を上げていないことから残響環境下における了解度向上において変調周波数成分を 8 dB 上げる必要がない可能性がある．

残響時間 5 s では各条件で了解度に明確な差が生まれていない先行研究 [7] より残響時間が長い時に clear speech においても了解度が上がっていないことから後部残響によるマスキングが大きすぎるとき，変調周波数成分を引き上げるだけでは了解度向上に至らないと推察される．しかし，残響時間 5 s の条件において図 4.6 で分散を確認したところ原音に比べて Gain control を施した音声は正答率の分散が大きく，原音に対してモーラ正答率が高くなる聴取者がいた．このことから残響時間が長いとき，変調周波数成分を引き上げた音声の了解度には個人差が生まれている可能性がある．

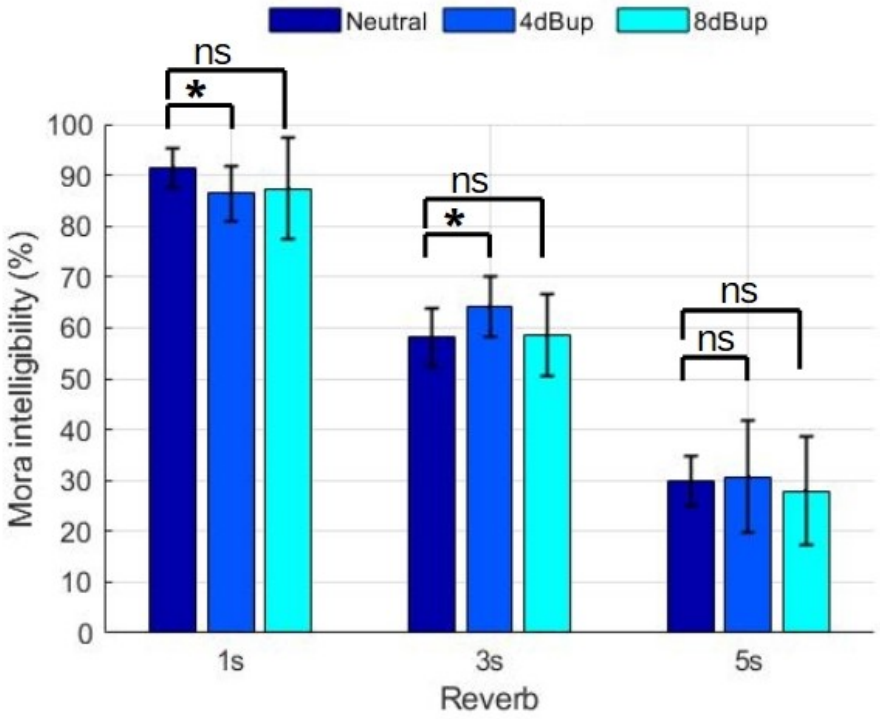

図 4.6: 聴取実験のモーラ別正解率 (*：p¿0.ns:有意差なし)

第5章　結論

5.1　明らかにしたこと

本研究では残響環境下でのヒトの発話変形を先行研究での得られた特徴と音声の変調スペクトルに着目した上で調査した．母音のＦ1，Ｆ2に着目し，変調スペクトルを分析した結果，ヒトは残響時間が長いときの発話において変調スペクトルの変調周数成分を増幅させていることが分かり，なかでも変調周波数10 Hz付近での増幅が顕著であったことから10 Hz付近に了解度向上の手掛かりがあることを示唆した．また，Ｆ2よりもＦ1の方が変調周波数成分が大きく増幅していることからＦ1の方が残響環境下での了解度向上に関係している可能性があることがわかった．

発話変形を分析することで得られた特徴量を用いて実験刺激を作成し，残響環境下での了解度調査を行った．この調査により，残響時間3sの残響環境において特定の変調周波数成分を引き上げることは了解度向上に有意であることを示した．また，ヒトは残響時間の長い残響環境での発話において変調周波数成分を引き上げる補償運動を行い，了解度向上を図っていることを明らかにした．

この研究の波及効果として，残響が存在し，音が聞き取り難い室内公共空間においても正確な情報を伝達することができるようになり，駅ホームなどでの災害時の避難誘導音声のような正確な情報が求められる状況にも役立てることができると考えられる．

5.2　残された課題

今回，母音のＦ1，Ｆ2両方の周波数帯域から出力されたパワーエンベロープに対して変調周波数成分の引き上げを行ったが，Ｆ1に該当する変調周波数成分のみを引き上げた音声とＦ2に該当する変調周波数成分のみを引き上げた音声を用いて了解度調査を行い，比較，検討することで了解度に大きく関わっている情報を特定する必要がある．また，先行研究で報告されている母音のＦ1，Ｆ2の成分と本研究で得られた変調周波数成分の両方を変化させた音声を合成し，了解度調査を行うことで了解度向上に最も有意である音響特徴量の検討する必要がある．

謝辞

本研究を進めるにあたり，ご指導して頂いた赤木教授に心から御礼申し上げます.
研究室会議において多くの助言を下さった鵜木教授に心から感謝致します.
研究計画を進めるうえで多く助言をしてくださり、本研究に必要なデータの提供をしてくださった脳情報通信融合研究センター研究技術員の久保 理恵子氏に心から感謝いたします.
日頃から助言をして頂き，ご協力してくださった研究室の皆様に心から感謝致します.
最後に本学での研究活動を支え，温かく見守ってくれた家族に心から感謝致します.

参考文献

[1] Cooke, M., King, S., Garnier, M., Aubanel, V., “The listening talker: A review of human and algorithmic context-induced modifications of speech, ”Comput. Speech Lang, vol. 28, no. 2, pp. 543–571, 2014.

[2] Lombard, E., “Le signe de I ’ elevation de la voix, Ann. Mal. De L, ”Oreille et du Larynx, vol. 37, pp. 101–119, 1911.

[3] T.V. Ngo, R. Kubo, D. Morikawa, and M. Akagi, “Acoustical analyses of tendencies of intelligibility in lombard speech with different background noise levels, ”Journal of Signal Processing, vol. 21, no. 4, pp.171—174, 2017.

[4] A.K. Nabelek, T.R. Letowski, F.M. Tucker, “Reverberant overlap- and self-masking in consonant identfication, ” J. Acoust. Soc. Am., vol. 86, no. 4, pp. 1259—1265, 1989.

[5] 程島奈緒, 荒井隆行, 栗栖清浩, “雑音・残響下の発話による音声の明瞭度改善,” 日本音響学会 2010 年秋季研究発表会講演論文集, pp.521–524, 2010.

[6] 程島奈緒, 荒井隆行, 栗栖清浩, “「はっきり」と発話した音声の明瞭度と聴覚印象評価～残響下を想定した発話の場合～,” 日本音響学会 2008 年秋季研究発表会講演論文集, pp.345–346, 2008.

[7] N. Hodoshima, T. Arai, and K. Kurisu, “Effects of training, style, and rate of speaking on s peech perception of young people in reverberation, ” Proc. Acoustics 08, pp. 2393–2397, 2008.

[8] 荒井隆行, 木下慶介, 程島奈緒, 楠本亜希子, 喜田村朋子, “音声の定常部抑圧処理の残響に対する効果,” 日本音響学会研究発表会講演論文集, vol. 1, pp. 449–450, 2001.

[9] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto and T. Kitamura, ”effects on suppressing steady-state portions of speech on intelligibility in reverberant environments” Acoust. Sci. & Tech., 23, pp. 229–232 2002.

[10] 辻美咲, 荒井隆行, 安啓一, “残響環境における音声明瞭度改善を目的とした子音強調・母音抑圧による前処理,” 日本音響学会誌, vol. 69, no. 4, pp. 179–183, 2013.

[11] 久保理恵子, 森川大輔, 赤木正人, “発話時の残響時間の違いが 残響下での音声了解度に与える影響,” 日本音響学会 2017 年秋季研究発表会講演論文集, pp.369–370, 2017.

[12] 久保理恵子, 赤木正人, “発話時の残響時間によるフォルマント周波数の変化と残響下における了解度, ” IEICE technical report, vol. 117, no. 515pp.39–44, 2018.

[13] R. M. Uchanski, D. B. Pisoni, R. E. Remez, “The Handbook of Speech Perception(Clear speech),” Blackwell Publishing, chapter9, 2008.

[14] Schroeder, M. R., “Modulation transfer functions: definition and measurement,” Acustica, Vol. 49, pp. 179–182, 1981.

[15] Atlas, L., Greenberg, S., and Hermansky, H., “The Modulation Spectrum and Its Application to Speech Science and Technology,” Interspeech Tutorial, Antwerp, Belgium, 2007.

[16] M. Komatsu, T. Arai, ”Modulation Spectrum and Rhythmic Units of Japanese,” Journal of the Phonetic Society of Japan, vol.13, no.3, pp85–99, 2009.

[17] 近藤公久, 天野成昭, “親密度別単語了解度試験用音声データセット 2007 (FW07),” NTT 音声資源コンソーシアム, 2007.

[18] 桑原尚尾, 匂坂芳典, 武田一哉, 阿部匡伸, ”研究用 ATR 日本語音声データベースの作成 (別冊Ⅱ 不特定話者テキスト

) ,” ATR Technical Report, pp.26–47, 1989.

[19] L. Milic, “Multirate Filtering for Digital Signal Processing: MATLAB Applications,” University of Belgrade, 2009.

[20] Houtgast, T. and Steeneken, H. J. M., “The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility,” Acustica., vol. 28, pp. 66–73, 1973. 98

[21] R. Drullman, J. M. Festen, and R. Plomp, “Effect of reducing slow temporal modulations on speech reception, ” J. Acoust. Sco. Am., vol. 95, no. 5, pp. 2670–2680, 1994.

[22] 小林まおり, 鵜木祐史, 赤木正人, “了解性における音源の変調スペクトルと音環境の変調伝達関数の関係,” 日本音響学会研究資料, vol.47, no.5, pp.309–314, 2017.