

Title	騒音環境下における車室内発話音声の分析とその合成に関する研究
Author(s)	竹山, 佳成
Citation	
Issue Date	2006-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1994
Rights	
Description	Supervisor: 赤木 正人【鷗木祐史】, 情報科学研究科, 修士

修 士 論 文

騒音環境下における
車室内発話音声の分析とその合成に関する研究

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

竹山 佳成

2006 年 3 月

修 士 論 文

騒音環境下における 車室内発話音声の分析とその合成に関する研究

指導教官 赤木 正人 教授

審査委員主査 鵜木 祐史 助教授

審査委員 党 建武 教授

審査委員 小谷 一孔 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

410075 竹山 佳成

提出年月: 2006 年 2 月

概 要

現在，車室内音声認識システムへの期待が高まっているが，車室内ではロンバード効果等の発話歪みが発生し，車室内音声認識システムの精度を低下させる一因となっている．HMM を用いた音声認識では発話歪みに対応した音響モデルを学習するために車室内で収録された巨大な音声データベースが必要となるが，現在の音声データベースのほとんどは雑音下で収録されたものではなく，車室内発話音声サンプルの数は多くない．原因として車室内での様々な状態における音声収録が困難なことが挙げられる．

本研究では車室内音声認識システムの精度を改善するため，音響モデル学習用にクリーンな音声から車室内発話音声を合成する手法について述べる．自動車内を特徴付ける要因として自動車の速度に着目し，自動車の速度を関数とした音響的特徴量の変換関数を構築する．調査した特徴量は母音部における基本周波数，パワー，フォルマント周波数（第一，第二，第三），スペクトル傾斜，音韻長である．クリーン音声に対しこれらの特徴量について変換を行うことで車室内発話音声の合成を行い，得られた音声が音響モデルの学習に役立つかどうかの評価を行った．その結果，速度が上がるにつれてクリーン音声より合成音声の方が音響モデル学習において有用である，との結果を得た．

目次

第1章	序論	1
1.1	はじめに	1
1.2	本研究の背景	1
1.2.1	音声認識の基本原理	1
1.2.2	ロンバード効果	2
1.3	本研究の目的	3
1.4	本論文の構成	3
第2章	車室内発話音声合成システムの構築	5
2.1	はじめに	5
2.2	車室内発話音声合成システムの目的	5
2.3	提案手法	6
第3章	車室内発話音声の収録	8
3.1	目的	8
3.2	車室内発話音声の収録	8
第4章	車室内発話音声の分析	9
4.1	目的	9
4.2	音響的特徴量の分析方法	9
4.3	分析結果	12
4.4	考察	20
4.4.1	基本周波数について	20
4.4.2	パワーについて	20
4.4.3	フォルマント周波数について	20
4.4.4	スペクトル傾斜について	20
4.4.5	音韻長について	21
4.4.6	全体の考察	21
第5章	車室内発話音声の合成	22
5.1	システムの仕様	22
5.2	特徴量の変化量	22

5.3	合成音声の主観的な聴覚印象	25
第 6 章	合成音声の評価	26
6.1	はじめに	26
6.2	評価方法	26
6.3	測定結果	27
6.3.1	クリーン音声に対する速度ごとの結果	27
6.3.2	合成音声に対する速度ごとの結果	27
6.4	考察	28
第 7 章	全体考察	32
第 8 章	結論	34
8.1	本論文で明らかにされたことの要約	34
8.2	今後の課題	34
8.2.1	合成音声を用いた音響モデルの学習	34
8.2.2	子音に関する制御	34

図 目 次

2.1	車室内発話音声合成システムの概要	7
4.1	基本周波数の変化の例 F_0	11
4.2	クリーン音声を基準とした F_0 の変化量	11
4.3	テストコースでの基本周波数の変化	13
4.4	防音室での基本周波数の変化	13
4.5	テストコースでのパワーの変化	14
4.6	防音室でのパワーの変化	14
4.7	テストコースでの第一フォルマントの変化	15
4.8	防音室での第一フォルマントの変化	15
4.9	テストコースでの第二フォルマントの変化	16
4.10	防音室での第二フォルマントの変化	16
4.11	テストコースでの第三フォルマントの変化	17
4.12	防音室での第三フォルマントの変化	17
4.13	テストコースでのスペクトル傾斜の変化	18
4.14	防音室でのスペクトル傾斜の変化	18
4.15	テストコースでの音韻長の変化	19
4.16	防音室での音韻長の変化	19
5.1	車室内発話音声合成処理の流れ	23
6.1	クリーン時と防音室の発話におけるマハラノビス距離	29
6.2	クリーン時とテストコース走行時の発話におけるマハラノビス距離	29
6.3	合成音声と防音室の発話におけるマハラノビス距離	30
6.4	合成音声とテストコース走行時の発話におけるマハラノビス距離	30
6.5	マハラノビス距離の変化の平均	31

表 目 次

5.1	車室内発話音声合成関数 <code>clean2car</code> に与える引数の詳細	24
5.2	クリーン音声を基準とした変化量 $D(v)$ ($av^2 + bv + c$)	24
6.1	音響分析条件詳細	27

第1章 序論

1.1 はじめに

自動車の運転においてカーナビやエアコンを音声によって操作することで、ドライバーはハンドルから手を、道から目を逸らさないで済み、簡単に機器を操作することが可能となる。このため、車室内音声認識技術への期待が高まっている。しかし、防音室などの環境と違い自動車走行中等の実環境においてはエンジン音や風切り音などの走行雑音が混入してしまう問題点がある。また、仮に混入した走行雑音を完全に除去できたとしても、雑音中では話者が無意識に普段より大きな声、高い声でしゃべってしまうロンバード効果 [1] の発生や、実際に運転しているというタスクを行いながら発話している、ということから音声に歪みが生じ、通常の状態とは異なる発話になっていると考えられる。これらの要因は車室内音声認識システムにおいて悪影響を与え、認識率の低下を招いている。

これらの問題は音声認識システムにおける音響モデルが車室内発話音声にうまく対応していないことが原因の一つとして挙げられる。近年の多くの音声認識システムで用いられている HMM (Hidden Markov Model) に基づいた音声認識システムの中では高精度な音響モデルの構築が認識率の向上につながるということが知られているが、そのためには大量の学習データをもって音響モデルを構築することが必要である [2]。そのため、車室内音声認識においては大量の車室内発話音声サンプルが必要となる。しかし多くの音声データベースは防音室などで収録されたものがほとんどであり、自動車走行中などの実環境においては音声サンプルの数が少ない。また、自動車走行中には自動車の速度、車種、路面の状況や天気などといった様々な状態が考えられ、あらゆる状態においての音声収録は困難である。

以上のことから車室内環境において生ずる発話歪みの実態を調査し、車室内発話において音声の中のどのような音響的特徴量が変化するか明らかにすることができれば、車室内音声認識の性能向上に寄与することが期待できる。

1.2 本研究の背景

1.2.1 音声認識の基本原則

音声認識とは音声波に含まれる言語情報を計算機によって抽出し、書き起こすことである。音声認識によって目や手足を使った作業中においても機械に情報を与えることがで

き，また，キーボードなどで手作業で文字を入力することに比べて容易であることなど，マシンインターフェースとして優れた利点があるとされる．

一般に音声認識の過程は確率モデルをもって表すことができる．入力音声の時系列パターンを $X = x_1, x_2, \dots, x_T$ のようにする．各 x_i はフレームごとに見た第 i フレーム音声の特徴ベクトルである．この中に単語系列 $W = w_1, w_2, \dots, w_n$ が含まれているとした場合，音声特徴量ベクトルの時系列 X から単語系列 W を推定することで認識を行う．これは $P(W|X)$ を最大にする単語系列 W を見つけることに相当する．ここでベイズの定理により，

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} \quad (1.1)$$

と定式化できる． $P(X)$ は入力音声自体の生起確率であるから単語系列 W に独立である．そのため， $P(X)$ は考慮せずに，次式を満たす W を見つければよい．

$$\arg \max_w P(X|W)P(W) \quad (1.2)$$

$P(X|W)$ は音響モデルとしてモデル化され， $P(W)$ 言語モデルとしてモデル化することができる．したがって音声認識の精度向上におけるポイントは音響モデルと言語モデルをいかに作るかによる．

1.2.2 ロンバード効果

雑音下における発話では雑音の影響から聴覚フィードバックが阻害され，話者が無意識に声を高く，大きくする，など普段の発話とはことなるしゃべり方をしてしまう．この現象はロンバード効果と呼ばれる．ロンバード効果によって発生様式の変化が起こると音声の特徴量の変化が起こり，その結果音声認識においては悪影響を与える．車室内音声認識においても走行雑音の影響からロンバード効果が発生し，認識率を低下させていると思われる．

ロンバード効果によって音声は次のような変化をすることが報告されている [4]．

- パワーの増大
- 基本周波数の上昇
- フォルマント周波数のシフト
- スペクトルの高域成分の増加

ロンバード効果による音声認識システムの性能低下は音声認識システムの音響モデルがロンバード効果によって変形された音声に適合していないため起こる．そこで，ロンバード効果発生時の発話様式の変化を分析し，認識時に補正を行うというアプローチが考えら

れる ([5], [6]) . しかし , ロンバード効果による認識率の低下は話者や音韻ごとによって異なることが報告されているため [7] , あらゆる場合における変化の補正は困難であると思われる . また , ある要因によって引き起こされる音声の変化は , 実環境など様々な要因が混在した場合 , 線形ではなく非線形に発生するとされる . このため , 適切な補正をすることが難しく , なかなか満足できる認識率を得ることができないのが現状である .

1.3 本研究の目的

本研究では車室内音声認識システムの性能向上のため , 車室内発話において生ずる発話歪みの実態を明らかにする . また , その結果を元にクリーンな音声を車室内発話音声に変換するシステムの構築を行う . 本システムを用いて多量の車室内発話音声を合成し , それを音響モデルの学習に用いることで車室内音声認識システムの性能向上が見込まれる .

車室内環境には様々なものが考えられる . 本研究では車室内環境を特徴付ける要因として自動車の速度に着目した . 車種や路面状況などは固定し , ある速度ごとに発話された音声サンプルを元に速度に応じて生ずる音響的特徴量の変化を分析し , そこから速度に応じた音響的特徴量の変換関数を求める . これにより 1 つのクリーン音声から任意の速度に応じた車室内発話音声を合成することが可能となる .

音響的特徴量の抽出および音響的特徴量の制御には音声分析変換合成法 STRAIGHT [8] を採用する . STRAIGHT によって個々の特徴量を分解し , それぞれを構築した変換関数をもって変換後 , 合成する . これにより特徴量を制御するだけでなく , 高品質な合成音声を得ることができると期待される .

1.4 本論文の構成

本論文の構成を以下に示す .

第 1 章では車室内音声認識システムの問題点および本研究の背景について説明し , それをもとに本研究の目的を明らかにした .

第 2 章では本研究において構築する車室内発話音声合成システムについてその目的 , 合成を行うための提案手法について述べる .

第 3 章では本研究において収録した車室内発話音声サンプルについて示す .

第 4 章では音響的特徴量の分析方法を示し , 3 章において収録した音声をもとに音響的特徴量の変化を分析する . 分析した特徴量は基本周波数 , パワー , フォルマント周波数 , スペクトル傾斜 , 音韻長の 5 種である .

第 5 章では 4 章において得られた結果を元に各音響的特徴量の変換関数の係数を決定し , 構築した車室内発話音声合成システムについて述べる . また , 合成した音声についての主観的な聴覚印象についても述べる . 第 6 章では本システムによって得られた車室内発話音声 HMM 音響モデルの学習に有用であるかを評価する . 評価には音声認識システ

ムにおいて特徴量ベクトルとして用いられている MFCC などのパラメータを用いた．

第7章では，全体の考察を行い，第8章にて本論文で得られた結果をまとめ，今後の課題を示す．

第2章 車室内発話音声合成システムの構築

2.1 はじめに

本章では本研究において構築する車室内発話音声合成システムの目的・概要について説明する．

2.2 車室内発話音声合成システムの目的

近年，音声認識システムの進歩によって雑音のない状態においては高い認識率を得ることが可能となった [3]．しかし，自動車走行中における車室内環境などの実環境における音声認識においては精度においてまだまだ問題がある．この原因として音声認識システムにおける音響モデルが雑音中の発話に適合していない，ということが挙げられる．近年の多くの音声認識システムは HMM に代表される統計的手法をもって認識を行っているが，高精度な認識を行うためには多量の音声サンプルが必要となる．しかし，既存の音声データベースのほとんどは防音室などの環境で収録されたものがほとんどであり，車室内環境などの実環境における音声データベースはほとんど存在しない．原因として，車室内環境における音声収録が非常に困難なことが挙げられる．また，車室内環境には様々な状態が考えられ，あらゆる状態における音声収録は不可能である．

そこで本研究では車室内音声認識システムの精度向上を目的とし，車室内発話音声において生ずる発話歪みの実態を明らかにする．本研究では車室内環境を特徴付ける要因として自動車の速度に着目した．車種や路面状況などは固定し，速度の変化によって生ずる音響的特徴量の変化の分析を行う．

また，本研究では得られた結果を元にクリーンな音声を任意の速度に応じた車室内発話音声に変形合成するシステムの構築を行う．クリーンな音声から車室内発話音声を合成することが可能になれば既存の音声データベースを車室内発話音声データベースとして活用するなど，多数の車室内発話音声を得ることが可能となる．これら合成音声をもって音響モデルの学習を行うことで車室内音声認識システムの性能向上が見込まれる．

2.3 提案手法

図 2.3 に車室内発話音声合成システムの概要図を示す．変換システムは，音響的特徴量の抽出を行う解析パート，速度に応じて特徴量を変換する変換パート，変換した特徴量をもとに合成を行う合成パートの 3 つから成る．図において，入力 x がクリーン音声で出力 y が合成した車室内発話音声である．入力されたクリーン音声は解析部において個々の音響的特徴量に分解される．これらの音響的特徴量は速度の関数である $H(v)$ によって車室内発話音声を模擬した合成音声 y の特徴量に変換される．合成部では変換された特徴量をもとに合成を行う．このようにしてクリーン音声から車室内発話音声合成される．音響的特徴量の制御および高品質な合成を行うために，本研究では音声分析合成系として STRAIGHT を採用した．

本研究では基本周波数，パワー，フォルマント周波数，スペクトル傾斜，音韻長の音響的特徴量について速度ごとに変化の傾向を分析し，これらを車室内発話音声のものに変換した．これらの特徴量はロンバード効果において変化するとみられる代表的な特徴量である．

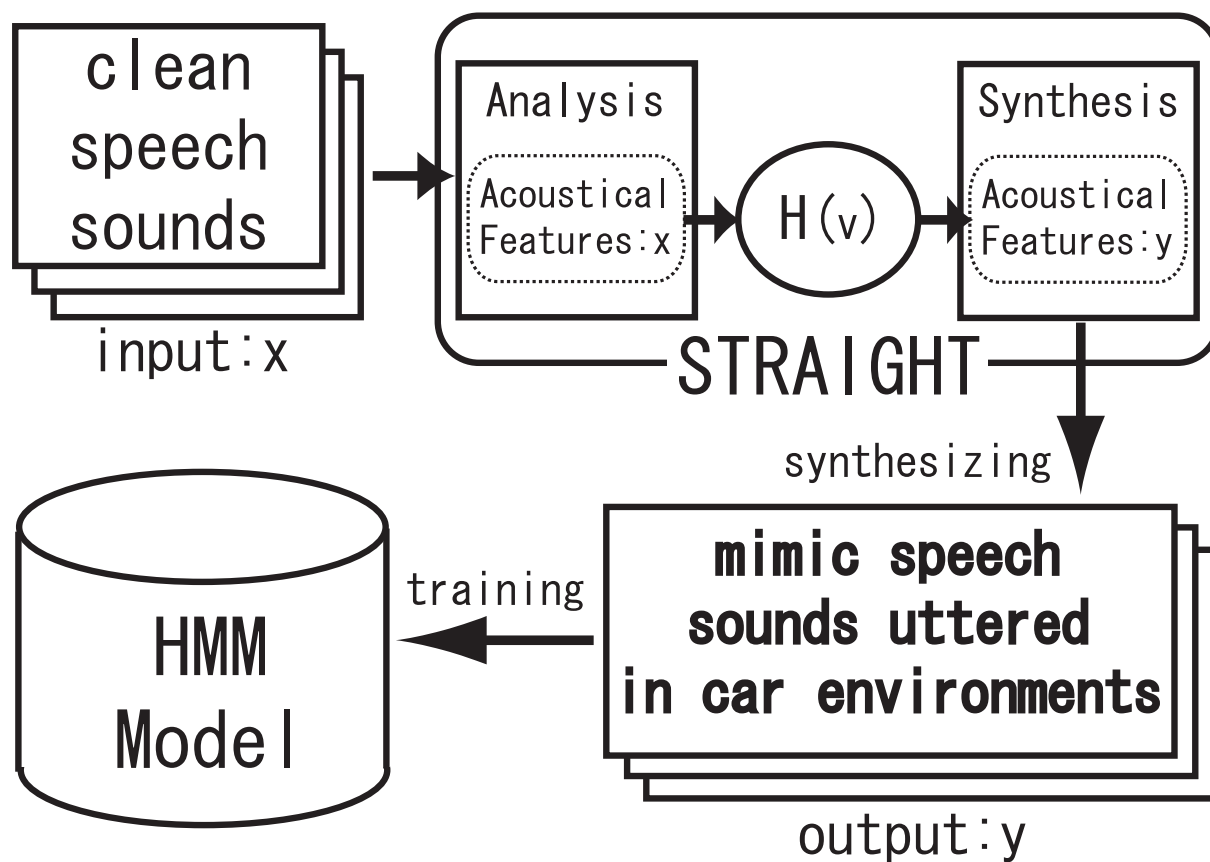


図 2.1: 車室内発話音声合成システムの概要

第3章 車室内発話音声の収録

3.1 目的

自動車の速度ごとに応じた車室内発話音声を合成するためには音声の音響的特徴量が自動車の速度とともにどのように変化しているか，を分析する必要がある．そのために分析対象としていくつかの車室内発話音声の収録を行った．速度に応じた特徴量の変化を見るため，ある特定の速度ごとに同一の単語の発話を収録した．

3.2 車室内発話音声の収録

分析対象とした車室内発話音声の詳細を以下に示す．

- 発話者
発話者は男性3名である．
- 収録環境
収録環境は実際にテストコース走行時に車室内の運転席において収録したものと，防音室においてヘッドフォンにより走行雑音を受聴しながらの発話を収録したものの2種類がある．これは実際に運転するというタスクを行いながら話す状態とヘッドフォンによる雑音受聴時の状態とで特徴量の変化に差が現れるかをみるためである．
- 雑音環境
雑音環境としては clean (防音室において雑音がない状態)，自動車の速度として 0 km/h (アイドリング時)，30 km/h，60 km/h，100 km/h，の5種類である．テストコース走行ではストレート部分でこれらの速度になった時に発話したものを収録し，防音室では同一テストコース走行時と同一車種による走行雑音を録音したものをヘッドフォンから呈示している．
- 収録単語
収録した単語は日本語5母音，および，ヤキトリ，ヤオヤ，ワスレモノ，ヒエコム，イク，ミハリ，の6単語である．これらの発話を各収録環境，雑音環境全部の場合において収録した．

第4章 車室内発話音声の分析

4.1 目的

本章では車室内発話音声について音響的特徴量の分析結果についてまとめ、その考察を行う。各音響的特徴量について速度ごとにどのような違いがあらわれるかを分析するとともに、各音韻の違いや、収録環境による違いについても調べる。

4.2 音響的特徴量の分析方法

自動車の速度に応じて音響的特徴量の変換を行うには、特徴量と自動車の速度との関連を調査する必要がある。クリーン音声と各速度における車室内発話音声を比較することにより特徴量の変化と速度との関連を分析を行い、特徴量の変換関数を構築した。

例として基本周波数の場合について変化の分析方法を説明する。図 4.1 は男性話者がクリーン環境で「あ」と発話した場合の基本周波数の時間軸上の推移である。水平軸が時間(ミリ秒)を、垂直軸が F_0 (Hz) である。図 4.1 における t_1 から t_2 のようにしゃべり始めとしゃべり終わりを除いた特徴量が安定している区間を定め、この区間において式 (1) を用いて音響的特徴量の平均値を計算した。

$$\overline{F_{0,v}} = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2-1} F_{0,v}[t] \quad (\text{Hz}) \quad (4.1)$$

このようにして 5 つの雑音環境 (clean, 0 km/h, 30 km/h, 60 km/h, 100 km/h) から 5 つの $\overline{F_{0,v}}$ が得られる。これらを元に式 (2) を用いてクリーンを基準とした各速度における変化量 $D_{F_0}(v)$ を計算した。

$$D_{F_0}(v) = \overline{F_{0,v}} - \overline{F_{0,\text{clean}}} \quad (\text{Hz}) \quad (4.2)$$

図 4.2 の上部に各速度における $D_{F_0}(v)$ の値を示す。これらの点を多項式で近似することにより、クリーンを基準とした速度に対する特徴量の変化量を表す関数を求めることが可能となる。図 4.2 の下部は上部の各点を最小誤差二乗法を用いて 2 次の多項式で近似したものである。他の特徴量についても同様に分析を行った。(パワー: $D_P(v)$, F_1 : $D_{F_1}(v)$, F_2 : $D_{F_2}(v)$, F_3 : $D_{F_3}(v)$, スペクトル傾斜: $D_T(v)$, 音韻長: $D_D(v)$)

クリーン音声の基本周波数を X_{F_0} , 車の速度を v とした場合 , 変換関数 Y_{F_0} は (3) のようになる .

$$Y_{F_0}(v) = X_{F_0} + D_{F_0}(v) \quad (\text{Hz}) \quad (4.3)$$

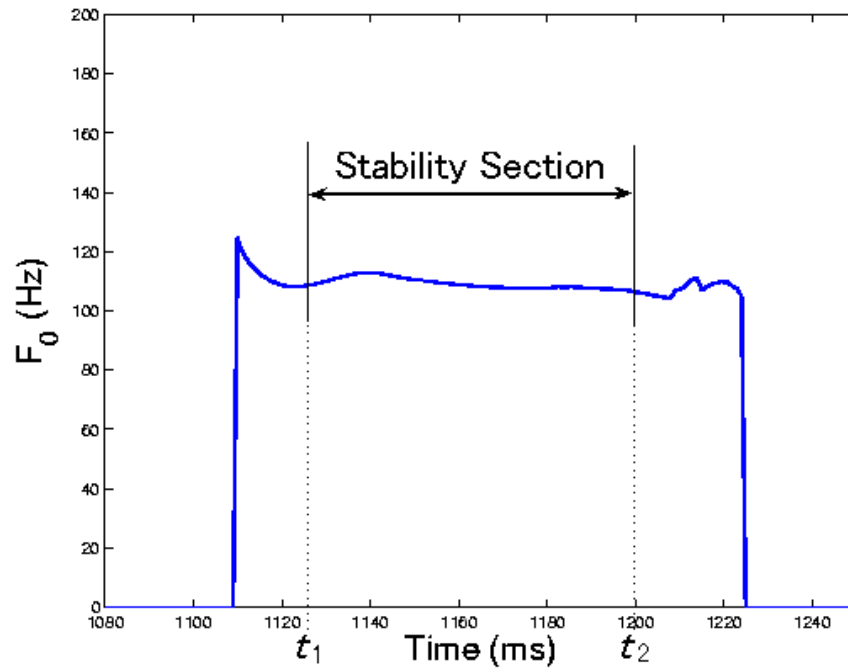


図 4.1: 基本周波数の変化の例 F_0

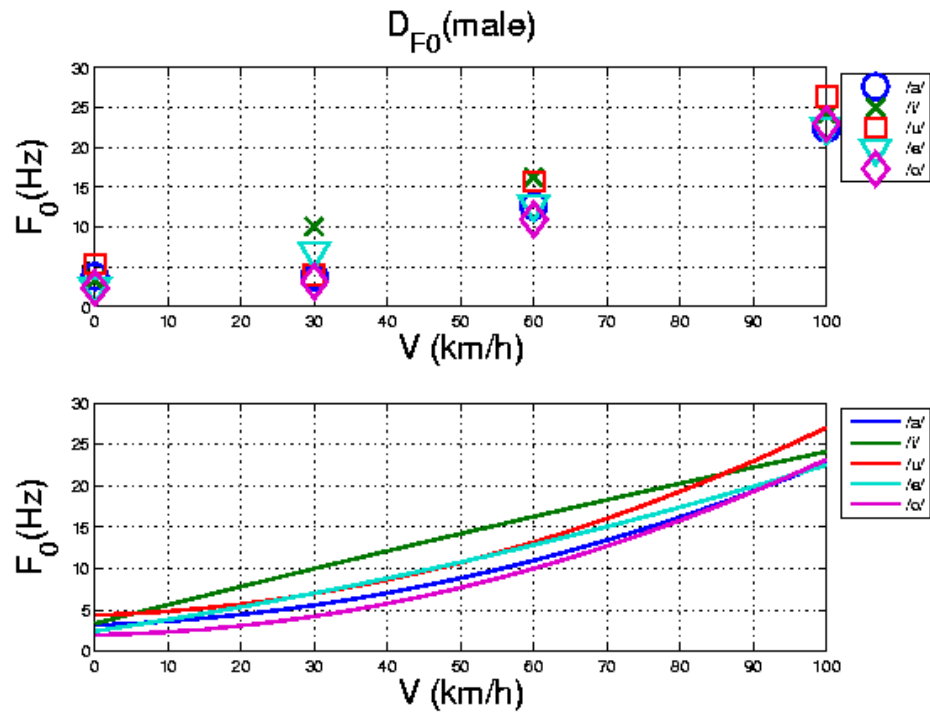


図 4.2: クリーン音声を基準とした F_0 の変化量

4.3 分析結果

図 4.3 から図 4.16 に各特徴量の変化を示す．横軸は自動車の速度 v (km/h)，縦軸はクリーン時を基準とした各特徴量の変化量である．点は収録音声から求めた各話者の変化の平均であり，曲線はそれらを最小誤差二乗法を用いて 2 次の多項式 ($av^2 + bv + c$) で近似したものである．基本周波数，パワー，フォルマント周波数，スペクトル傾斜については各母音ごとに色分けして表示し，音韻長については各モーラごとに色分けして表示している．

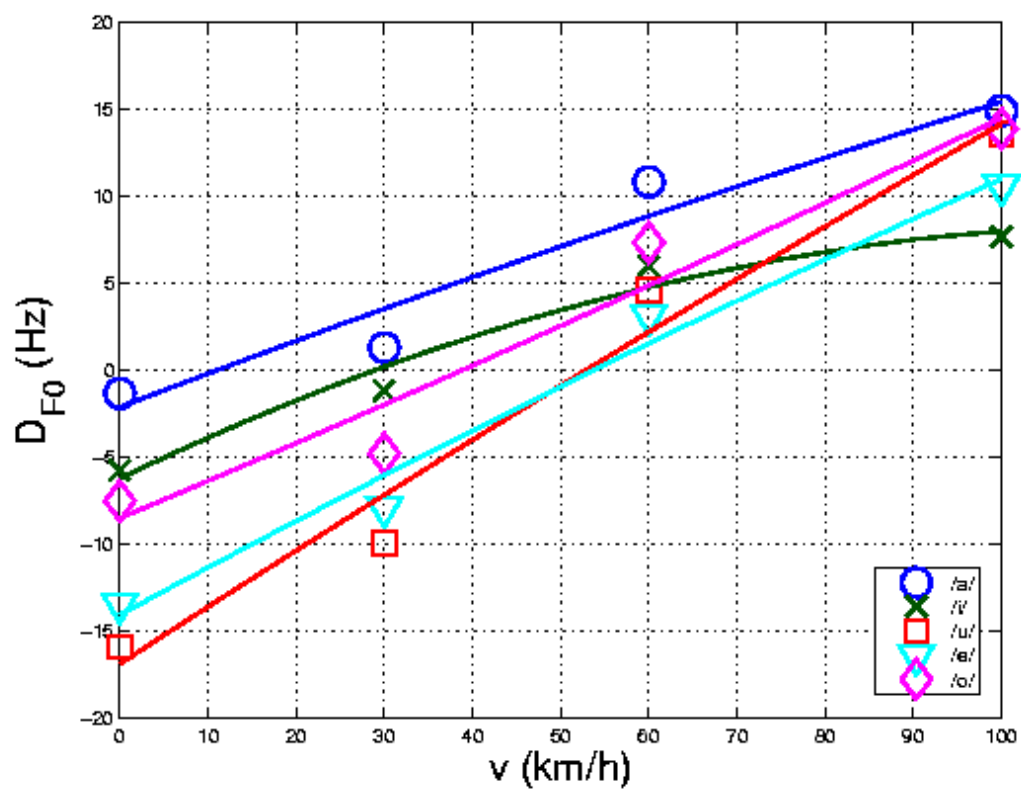


図 4.3: テストコースでの基本周波数の変化

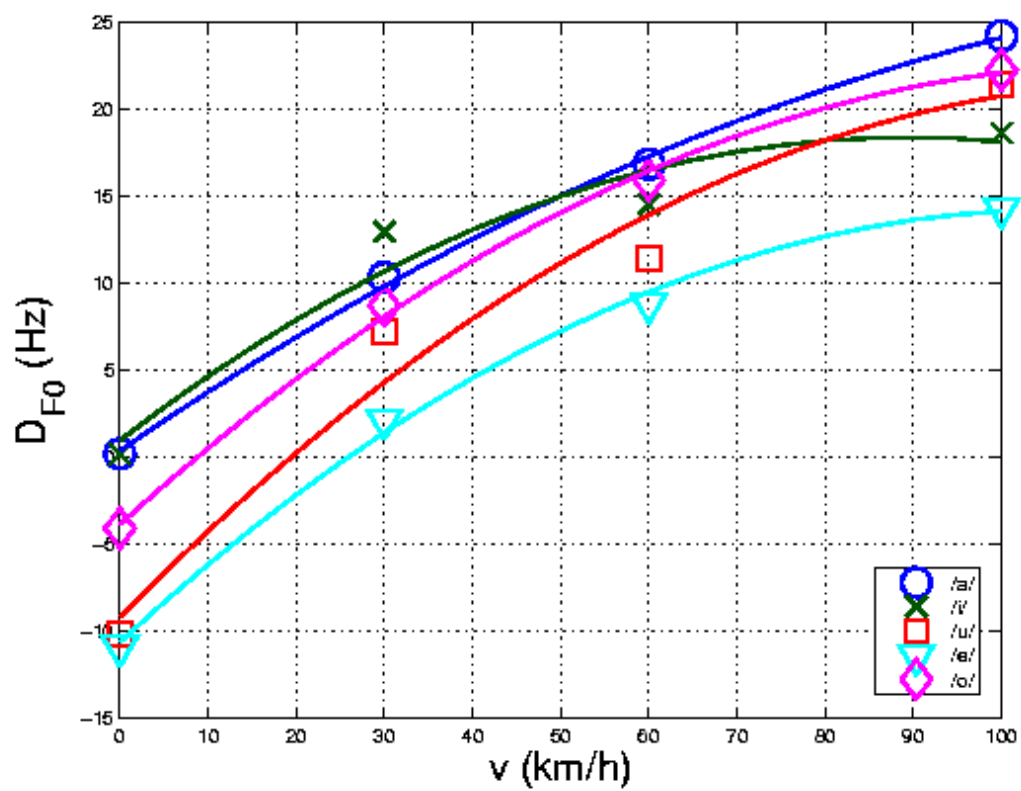


図 4.4: 防音室での基本周波数の変化

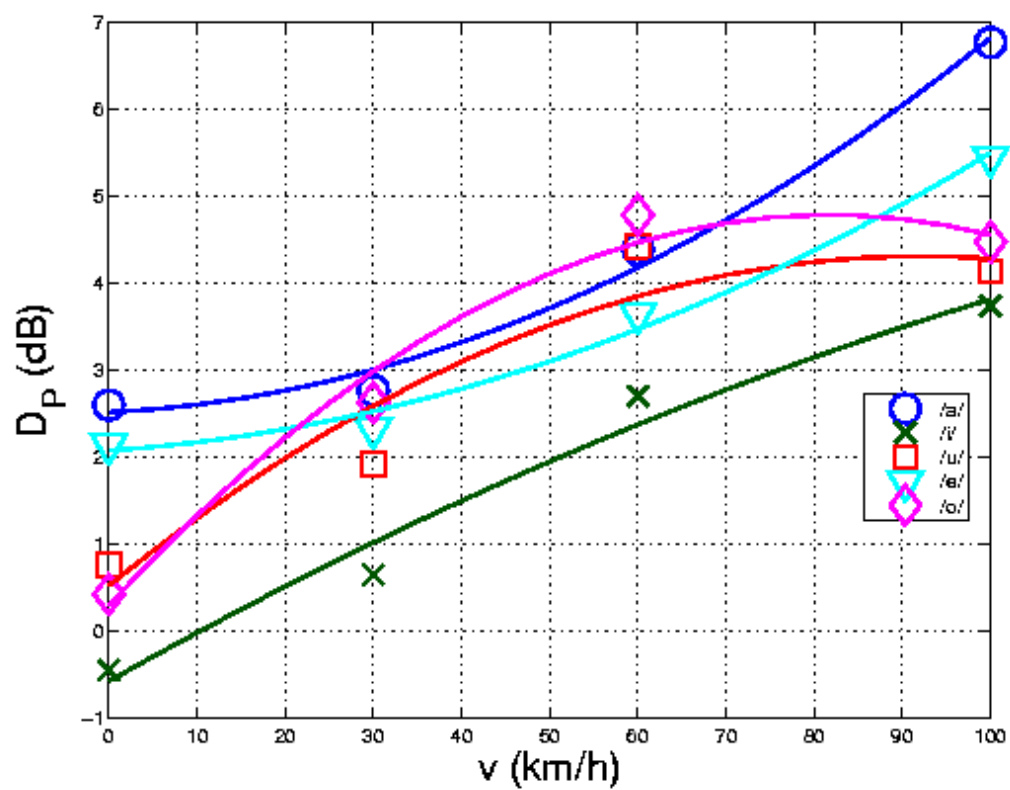


図 4.5: テストコースでのパワーの変化

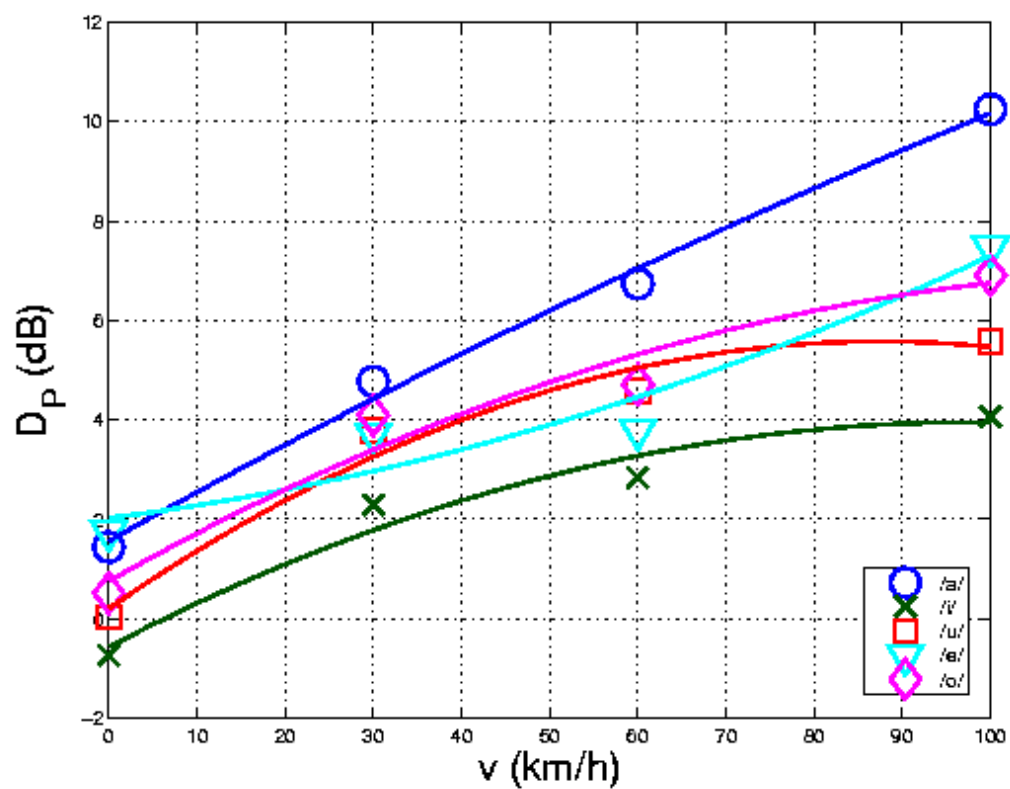


図 4.6: 防音室でのパワーの変化

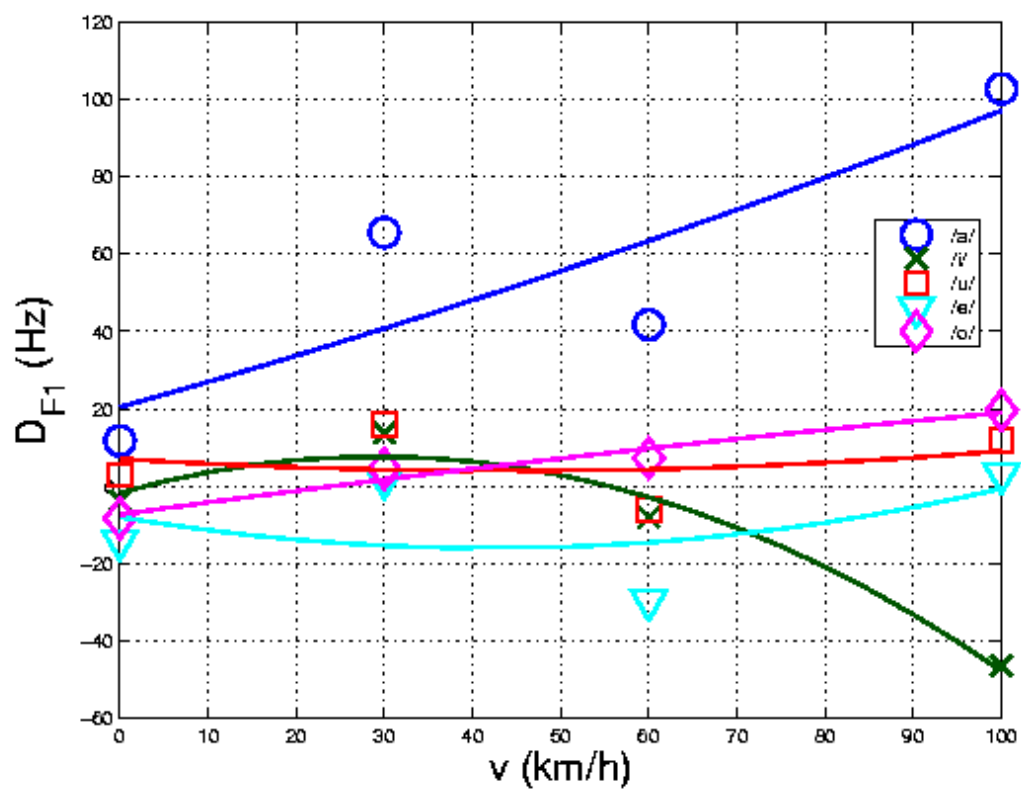


図 4.7: テストコースでの第一フォルマントの変化

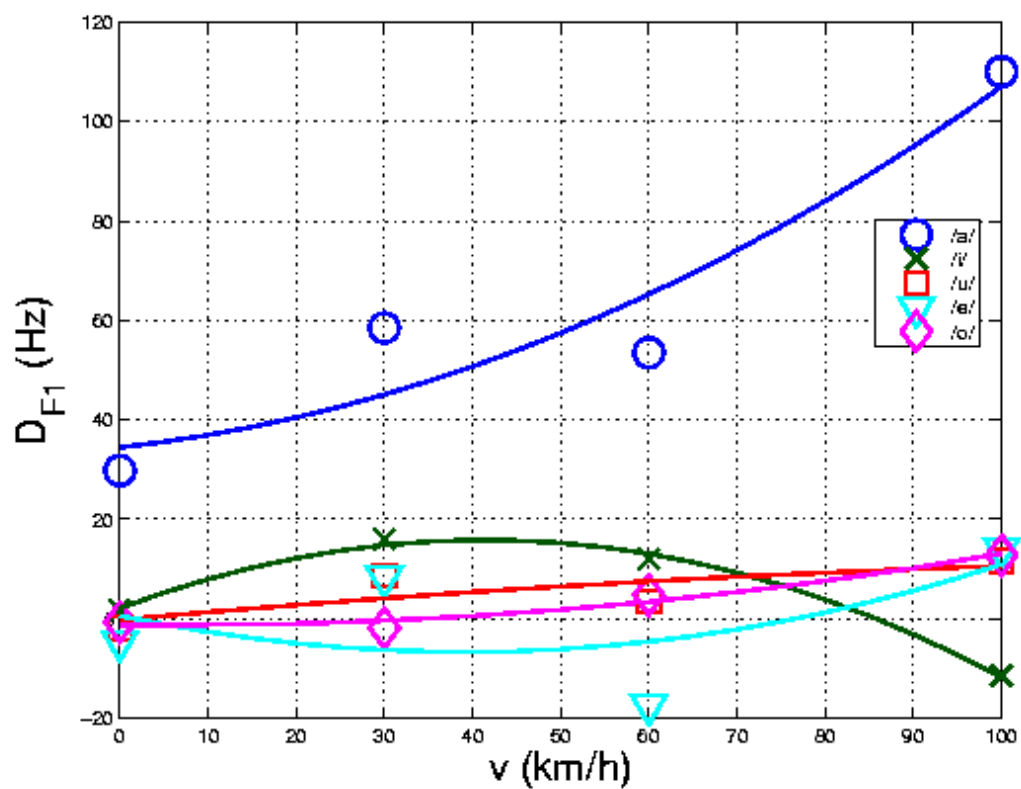


図 4.8: 防音室での第一フォルマントの変化

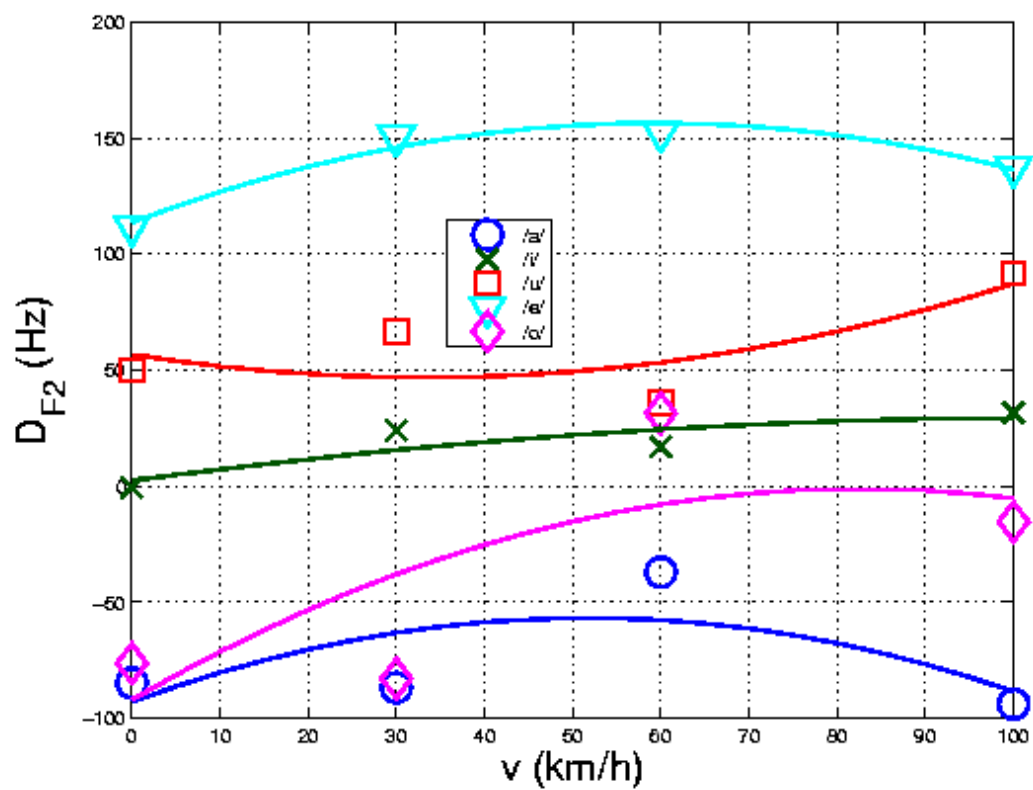


図 4.9: テストコースでの第二フォルマントの変化

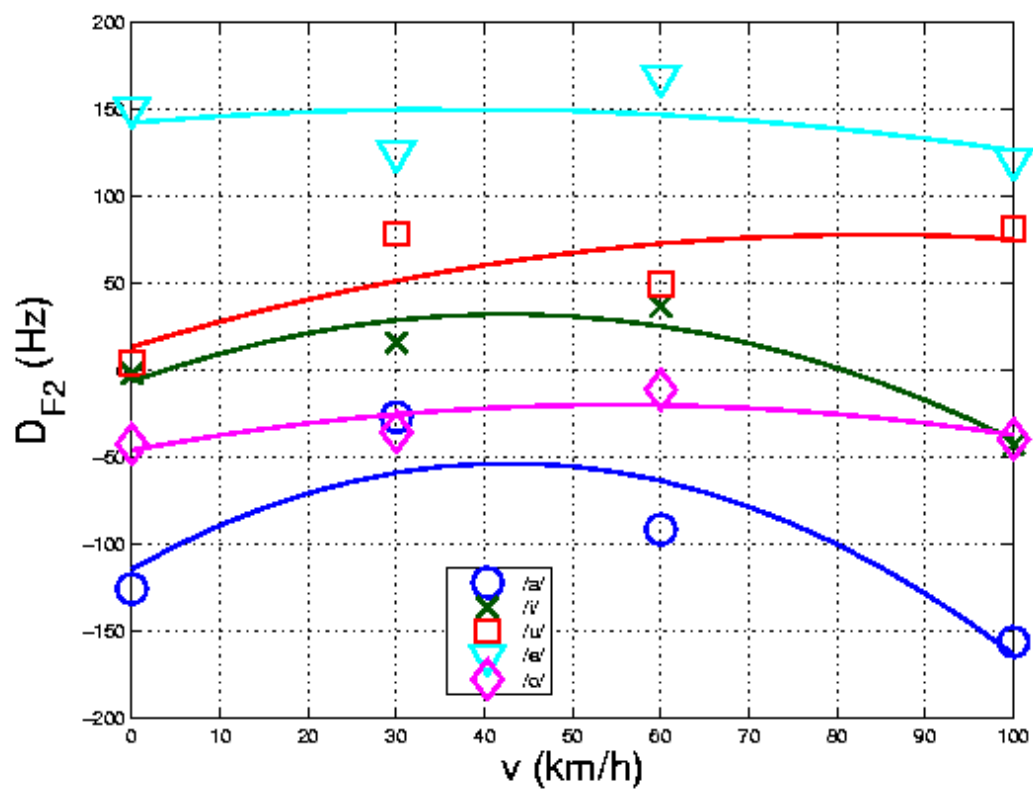


図 4.10: 防音室での第二フォルマントの変化

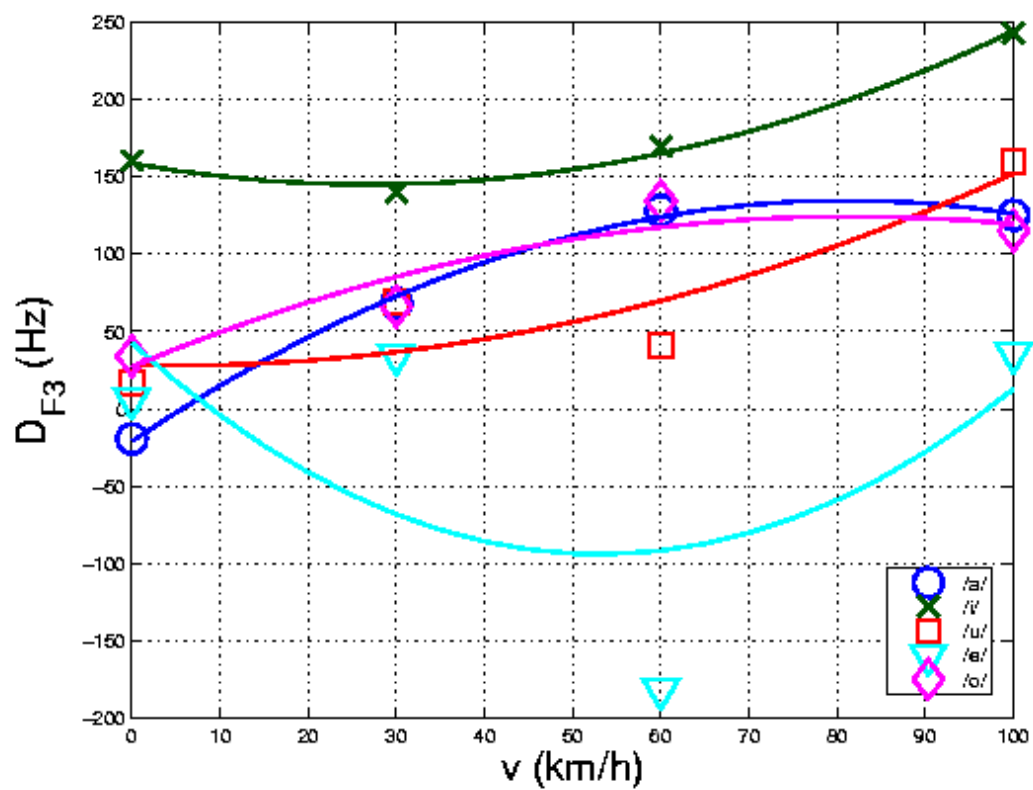


図 4.11: テストコースでの第三フォルマントの変化

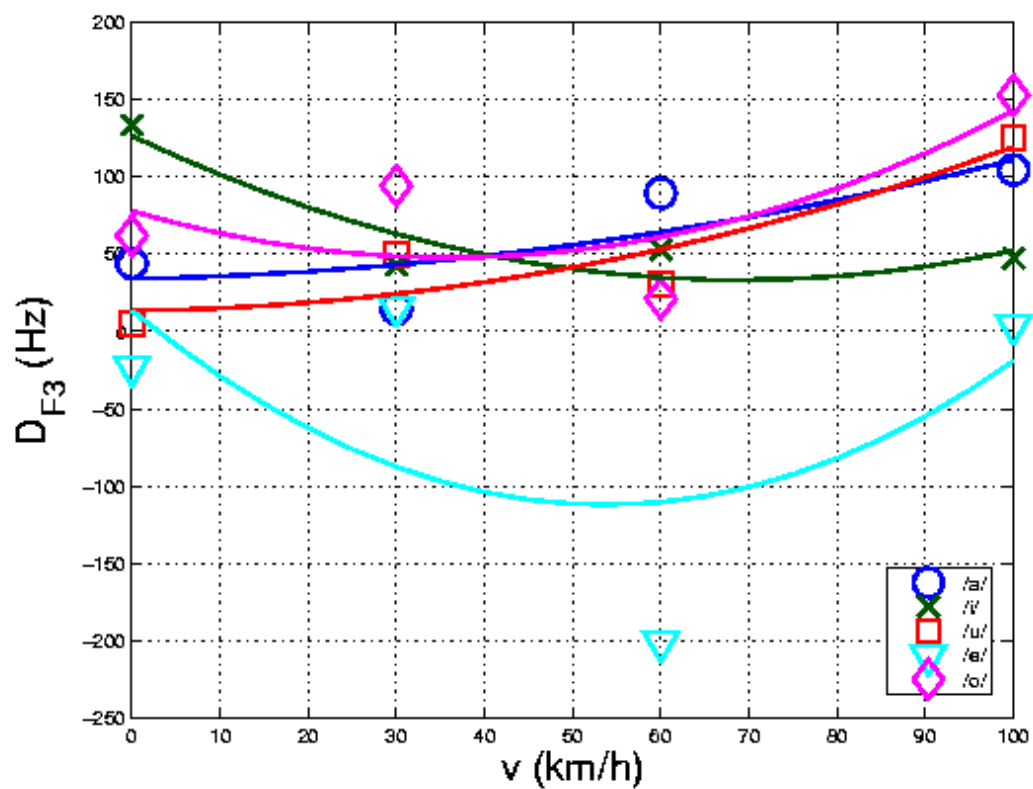


図 4.12: 防音室での第三フォルマントの変化

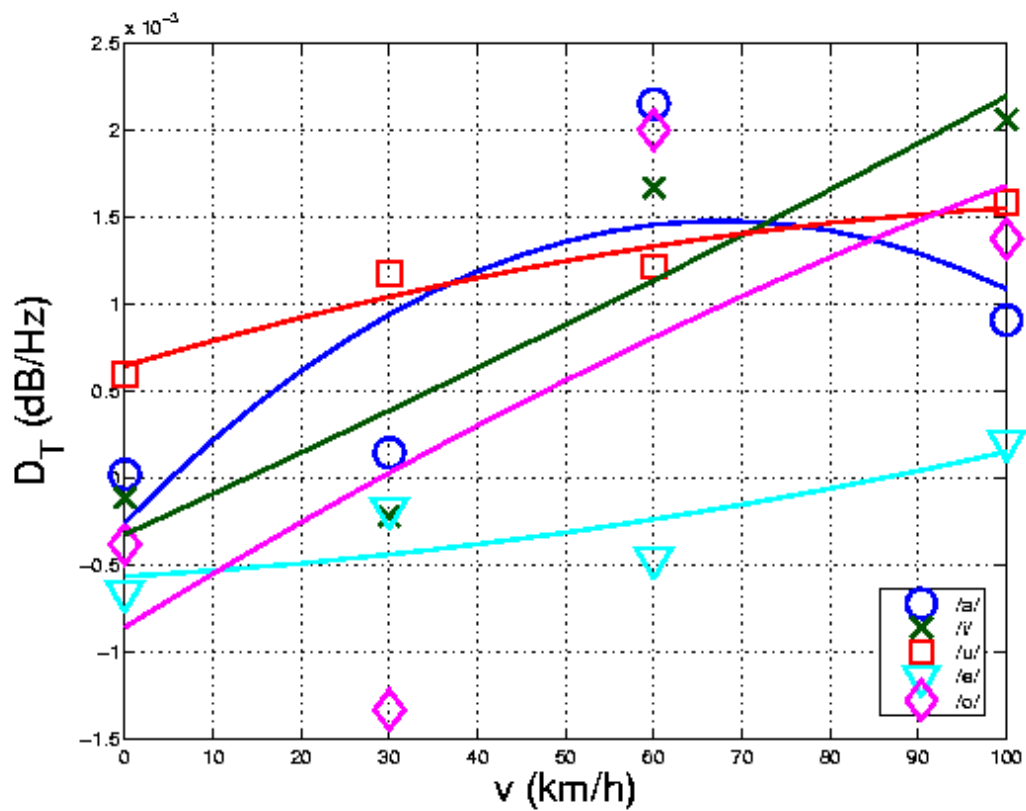


図 4.13: テストコースでのスペクトル傾斜の変化

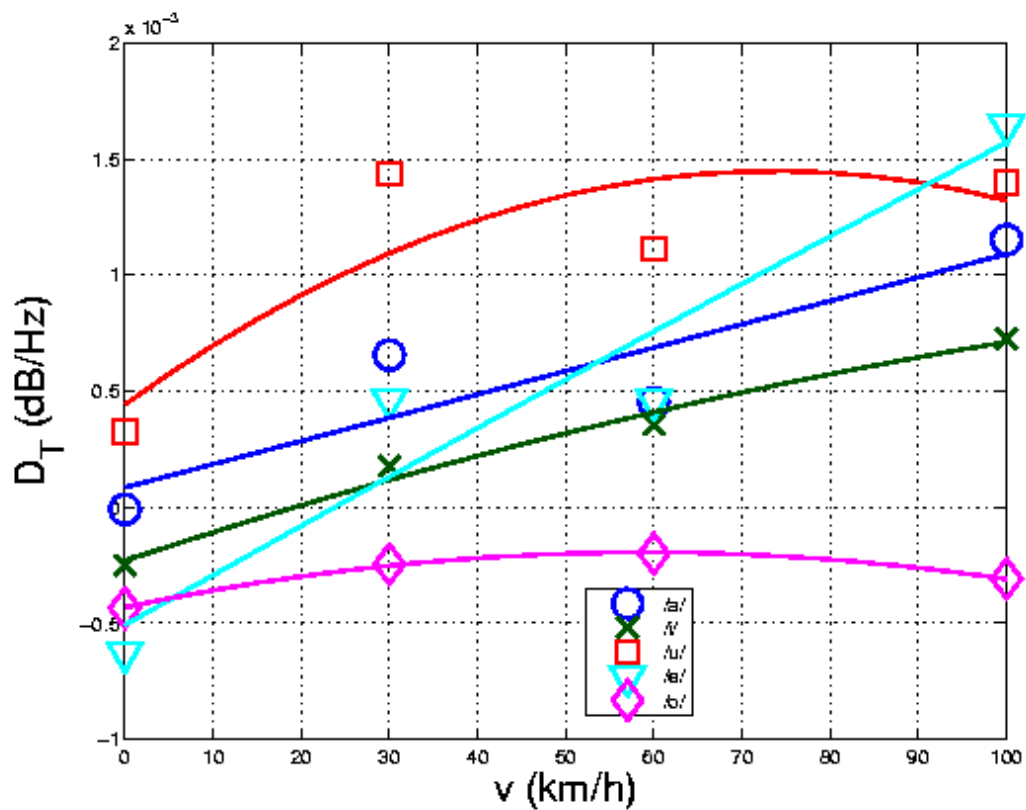


図 4.14: 防音室でのスペクトル傾斜の変化

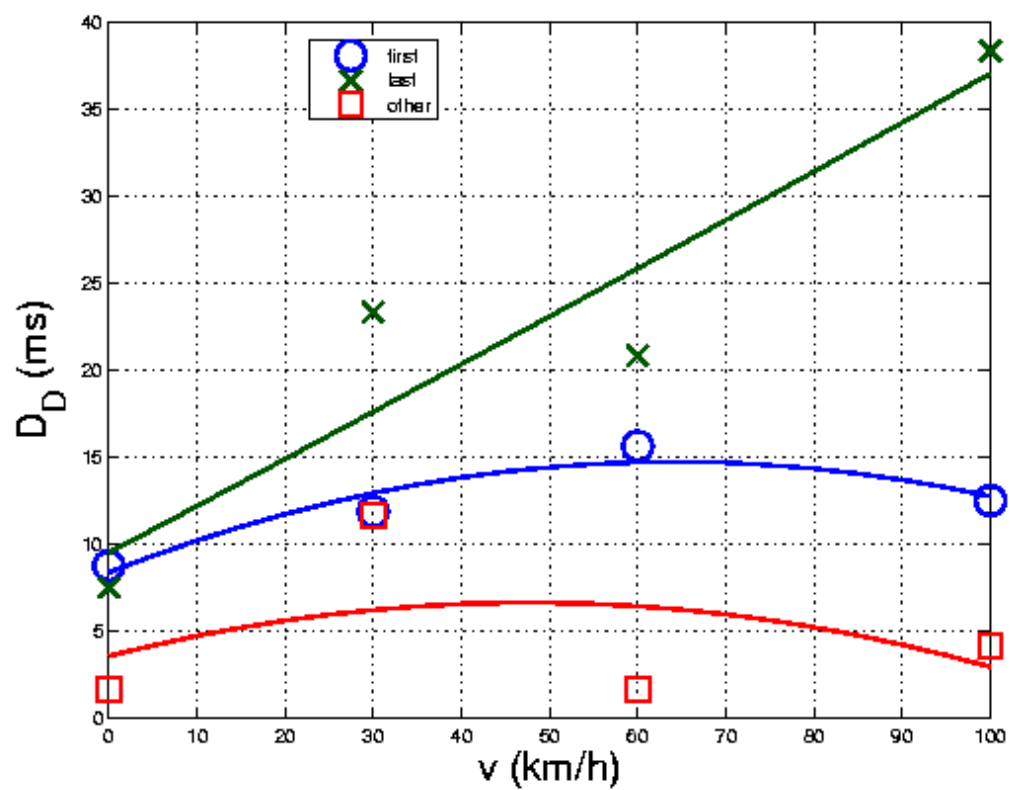


図 4.15: テストコースでの音韻長の変化

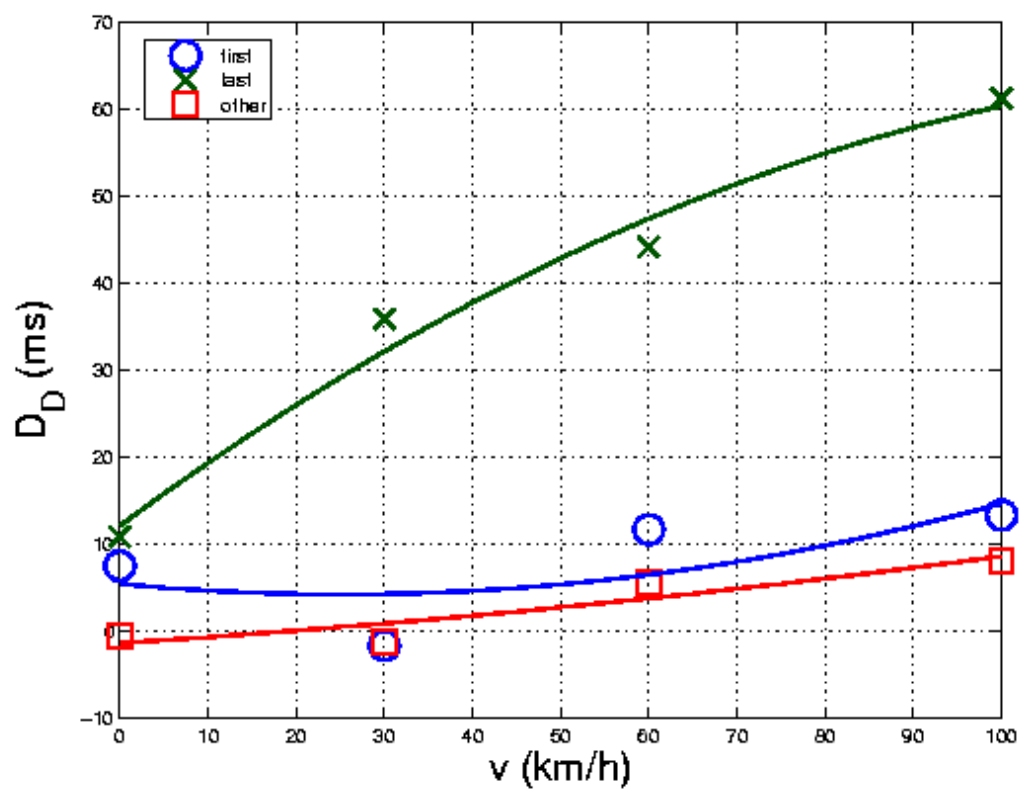


図 4.16: 防音室での音韻長の変化

4.4 考察

4.4.1 基本周波数について

図 4.3 にテストコース走行時，図 4.4 に防音室での基本周波数の変化を示す．各母音とも速度が上がるにつれて単調増加していることが見てとれる．また，テストコース，防音室で特に傾向に違いは見られない．これはロンバード効果が顕著に現れた結果であると考えられる．

4.4.2 パワーについて

図 4.5 にテストコース走行時，図 4.6 に防音室でのパワーの変化を示す．速度の上昇につれて単調増加していることが見てとれる．基本周波数の時と同様にロンバード効果が現れた結果と考えられる．また，各母音，各収録環境ともに目立った差は見られない．

4.4.3 フォルマント周波数について

図 4.7 から図 4.12 にフォルマント周波数の変化を示す．上部がテストコース走行時，下部が防音室である．収録環境の違いについては第三フォルマントについては若干ばらつきがあるが，だいたい同じような変化の傾向が見てとれる．母音ごとの変化については母音によっては値にばらつきがあり，また，変化の傾向も異なっている．ばらつきがあるものに関しては一部の話者，音韻によって値の変動が激しいものがあり，安定区間において値を正しく測定することができなかった可能性がある．また，一般的にロンバード効果においてはフォルマント周波数のシフトが起こるとされるが，低い周波数帯と高い周波数帯ではシフト方向が異なることが報告されている [9]．そのため，各音韻ごとや周波数帯でわけて変換関数を構築する必要があると思われる．

4.4.4 スペクトル傾斜について

図 4.13 にテストコース走行時，図 4.14 に防音室でのスペクトル傾斜の変化を示す．変化の傾向としては母音によっては少々差があるが，大体増加傾向にある．これは，ロンバード効果によってスペクトルの高域成分が増加し，傾斜がなめらかになっていくことを示している．収録環境の違いについてはテストコース走行では音韻によってはばらつきが激しいものがあるが，防音室のものはばらつきが少ない．テストコース走行時での収録においては発話音声に加えて少々走行雑音が混入してしまうため，このようにばらつきが出てしまった可能性がある．

4.4.5 音韻長について

図 4.15 にテストコース走行時，図 4.16 に防音室での音韻長の変化を示す．音韻長の変化についてはモーラの順番との間に強い関連が見られた．第一モーラと最終モーラが増加傾向にある．特に最終モーラについては強い増加傾向にあった．

4.4.6 全体の考察

以上のことから基本周波数およびパワーについては変化の傾向は母音に依存せず，そのため，全体を 1 つの変換関数でまとめてもよいものと思われる．フォルマント周波数およびスペクトル傾斜については各母音において変化の傾向が異なるものがあるため，母音ごとに変換関数を作るのがよいと思われる．また，音韻長ではモーラの順番でわけて変換関数を作るのがよいと思われる．また，収録環境についてはテストコース走行時，防音室で特別異なる傾向は見られなかった．

第5章 車室内発話音声の合成

5.1 システムの仕様

本研究において構築した車室内発話音声合成システムの仕様について説明する．本システムは MATLAB 上に動く関数として作成した．最終的に作成した関数は以下のようなものである．

```
function clean2car(inputFile, outputFile, v, s, option)
```

関数においてそれぞれの引数の詳細を表 5.1 に示す．

本システムが行う処理の流れを図 5.1 へ示す．入力として変換元のクリーン音声，自動車の速度，話者の性別，およびクリーン音声に対するセグメンテーションファイルを与える．セグメンテーションファイルとは音声に対し，含まれる音素の位置をミリ秒単位で記述しているテキストファイルである．その後，システムはセグメンテーションファイルを元に，Analysis part において母音区間に対して音響的特徴量を抽出する．抽出された特徴量は $H(v)$ にて速度に応じて変換される．変換は音韻長の増加，基本周波数の上昇，フォルマント周波数のシフト，スペクトル傾斜の増加，パワーの増加，の順番に行っている．convertDuration, convertF0, convertFormant, convertTilt, convertPower がそれぞれの処理を行う関数に相当する．これらの処理が終わった後，変換された特徴量は Synthesis part に渡され合成が行われる．

5.2 特徴量の変化量

前節において得られた各特徴量の変換関数をもとに車室内発話音声の合成を行う．クリーン時を基準とした特徴量の変化を表す関数 $(av^2 + bv + c)$ をテストコース走行時の分析結果をもとに表 5.2 のように定め，これらの関数を元に合成を行った．

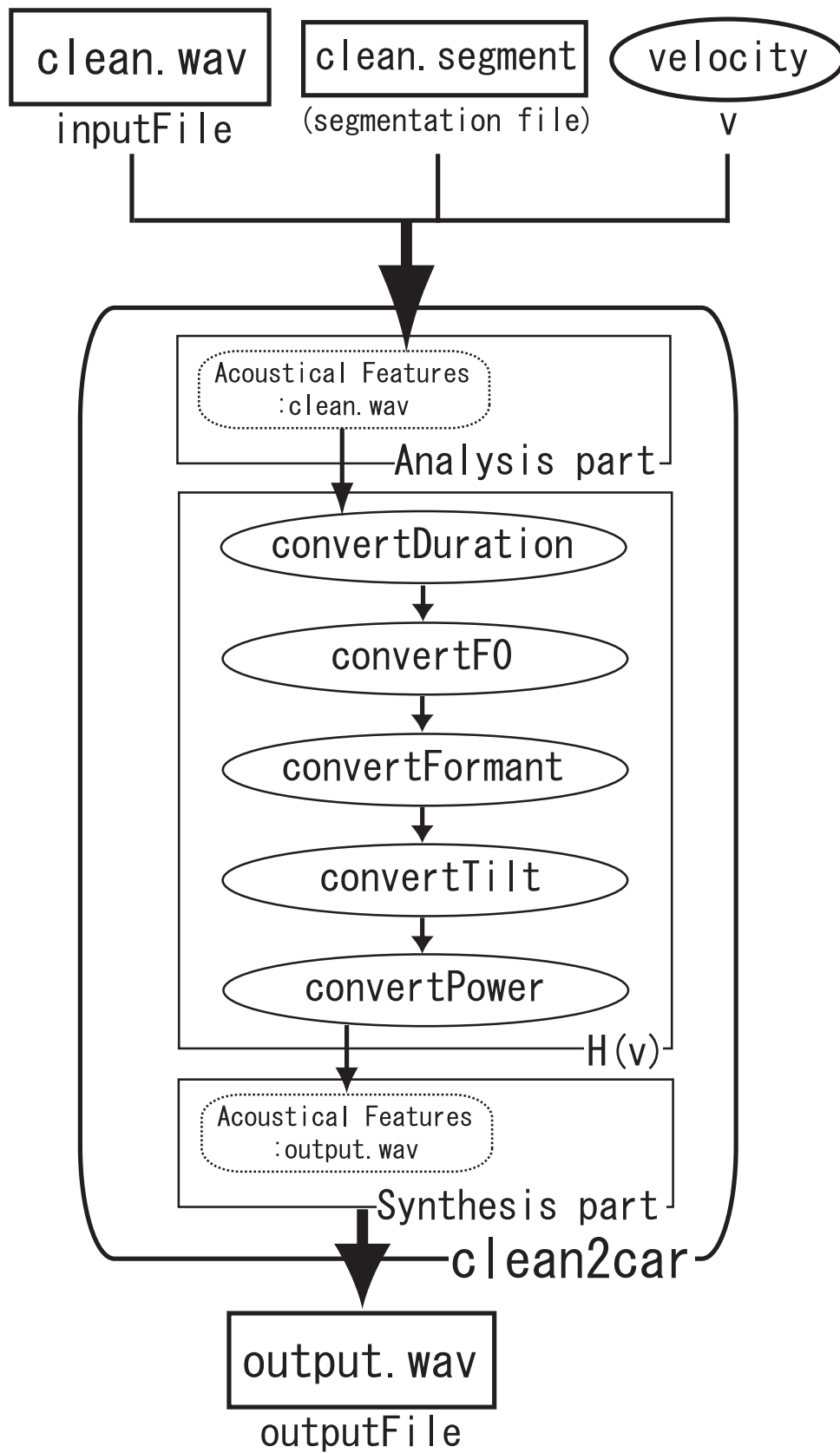


図 5.1: 車室内発話音声合成処理の流れ

表 5.1: 車室内発話音声合成関数 clean2car に与える引数の詳細

引数	役割
inputFile	合成元のクリーン音声のファイル名
outputFile	合成後に出力される音声のファイル名
v	自動車の速度 (km/h)
s	入力音声の話者の性別 (女性: 'f', 男性: 'm')
option	特徴量に対して変換を行うかどうかを指定する可変引数
'F0OFF'	基本周波数の変換を行わない
'powerOFF'	パワーの変換を行わない
'formantOFF'	フォルマント周波数の変換を行わない
'tiltOFF'	スペクトル傾斜の変換を行わない
'durationOFF'	音韻長の変換を行わない

表 5.2: クリーン音声を基準とした変化量 $D(v)$ ($av^2 + bv + c$)

特徴量	a	b	c
F_0 (Hz)	-2.9×10^{-4}	2.5×10^{-1}	-9.6
パワー (dB)	-1.2×10^{-4}	5.2×10^{-2}	9.7×10^{-1}
$F_1:/a/$ (Hz)	1.3×10^{-3}	6.4×10^{-1}	2.0×10
$F_1:/i/$ (Hz)	-1.1×10^{-2}	6.4×10^{-1}	-1.8
$F_1:/u/$ (Hz)	1.6×10^{-3}	-1.4×10^{-1}	7.1
$F_1:/e/$ (Hz)	4.6×10^{-3}	-3.8×10^{-1}	-8.0
$F_1:/o/$ (Hz)	-5.5×10^{-4}	3.1×10^{-1}	-7.4
$F_2:/a/$ (Hz)	-1.3×10^{-2}	1.4	-9.3×10
$F_2:/i/$ (Hz)	-2.3×10^{-3}	5.1×10^{-1}	2.3
$F_2:/u/$ (Hz)	9.0×10^{-3}	-6.0×10^{-1}	5.7×10
$F_2:/e/$ (Hz)	-1.2×10^{-2}	1.4	1.1×10^2
$F_2:/o/$ (Hz)	-1.3×10^{-2}	2.2	-9.2×10
$F_3:/a/$ (Hz)	-2.3×10^{-2}	3.8	-2.1×10
$F_3:/i/$ (Hz)	1.8×10^{-2}	-1.0	1.6×10^2
$F_3:/u/$ (Hz)	1.3×10^{-2}	-1.2×10^{-1}	2.8×10
$F_3:/e/$ (Hz)	4.8×10^{-2}	-5.1	4.2×10
$F_3:/o/$ (Hz)	-1.5×10^{-2}	2.4	2.7×10
傾斜:/a/ (dB/Hz)	-3.7×10^{-7}	5.1×10^{-5}	-2.6×10^{-4}
傾斜:/i/ (dB/Hz)	2.1×10^{-8}	2.3×10^{-5}	-3.2×10^{-4}
傾斜:/u/ (dB/Hz)	-6.0×10^{-8}	1.5×10^{-5}	6.4×10^{-4}
傾斜:/e/ (dB/Hz)	4.2×10^{-8}	2.9×10^{-6}	-5.6×10^{-4}
傾斜:/o/ (dB/Hz)	-6.0×10^{-8}	3.1×10^{-5}	-8.6×10^{-4}
音韻長:第一モーラ (ms)	-1.3×10^{-3}	1.3×10^{-1}	3.6
音韻長:最終モーラ (ms)	7.7×10^{-5}	2.3×10^{-1}	9.5
音韻長:その他 (ms)	-1.5×10^{-3}	2.0×10^{-1}	8.4

5.3 合成音声の主観的な聴覚印象

合成した音声と収録音声を聞き比べた際の主観的な聴覚印象について述べる．両方ともに速度の上昇につれてだんだんと声が高く，大きくなっていっているのが聞いてとれた．そのため，基本周波数およびパワーについての変換は正しく行われていると思われる．しかし，分析結果からフォルマント周波数およびスペクトル傾斜については音韻によっては値にばらつきがでたためか，声の質に関しては違和感を感じるものがあった．今回フォルマント周波数において第二，第三と高域部分にいくほど値の変動が激しかったため，実際の音声と合成音声ではこれらの値が離れており，声質に違いが出た可能性がある．しかし，合成音声を作る目的はそれが音響モデルの学習に利用できるかということが重要であり，音質が悪く感じるなど聴覚上異なって聞こえることは本質的な問題ではない．逆に言えば，一見聞いた感じでは収録音声とほぼ同様に聞こえたとしても，学習用データとしてはふさわしくない，という場合も考えられる．

第6章 合成音声の評価

6.1 はじめに

本章では構築したシステムを用いて合成した車室内発話音声と音声認識システムの音響モデル学習用として有用であるかの評価を行う。

6.2 評価方法

音声認識システムにおいては音声の特徴量ベクトルをもつて認識を行っている。そのため、ある速度において収録した車室内発話音声と合成した音声の特徴量ベクトルを調べ、両者の分布が近いほど音響モデルの学習用データにふさわしいものであると考えられる。そこで収録音声と合成音声についてそれぞれの特徴量分布の距離を調べることで評価を行う。

使用する特徴量ベクトルとしては MFCC (12 次元), Δ MFCC(12 次元), Δ Power (1 次元) を HTK (Hidden Markov Model Toolkit) [11] によって取得した。詳細な音響分析条件を表 6.1 に示す。

分布の距離を測るためにはマハラノビス距離を用いた [10]。音声 i と j におけるマハラノビス距離 D_{ij} を (1) のように定めた。

$$D_{ij} = \sqrt{\frac{K \sum_{k=1}^K (\mu_{ik} - \mu_{jk})^2}{\sum_{k=1}^K \sigma_{ik}^2 + \sum_{k=1}^K \sigma_{jk}^2}} \quad (6.1)$$

K は MFCC ベクトルの次元数 ($K = 25$) である。 μ_{ik} および σ_{jk} はそれぞれ音声 i における平均および分散特徴量ベクトルの k 次元目の要素である。

表 6.1: 音響分析条件詳細

サンプリング周波数	16 kHz
プリエンファシス	0.97
分析窓	Hamming 窓
分析窓長	25 ms
窓間隔	10 ms
特徴パラメタ	MFCC (12 次) + Δ MFCC (12 次) + Δ パワー (計 25 次)
周波数分析	等メル間隔フィルタバンク
フィルタバンク	24 チャンネル
CMS	発話単位

6.3 測定結果

男性話者 1 名に対し、各単語について収録音声と合成した音声とでマハラノビス距離の測定を行った。

6.3.1 クリーン音声に対する速度ごとの結果

図 6.1 および図 6.2 に i にクリーン音声固定、 j に横軸で示す速度において収録した音声を与えた時の D_{ij} の結果を示す。上部が防音室での発話であり、下部がテストコース走行時の発話である。どちらもだいたい速度の上昇につれて D_{ij} の値が増加していつているのが見てとれる。これは速度が上がるにつれ、クリーン音声と収録音声とでは特徴量ベクトル分布間の距離が離れていつてしまっていることを示している。また、防音室時の発話に比べ、テストコース走行時は値の増加が激しい。これはテストコース走行時の収録においてはいくらかの走行雑音が混入してしまっているためかと考えられる。走行雑音は速度が上がるにつれて増加するため、0 km/h では割と近い位置にあるものが速度の上昇につれて離れていつたと思われる。

6.3.2 合成音声に対する速度ごとの結果

図 6.3 および図 6.4 に i に合成した車室内発話音声、 j に収録した音声を与えた時の D_{ij} の結果を示す。それぞれ横軸で示す速度について、 i は表 5.2 に示した係数を用いて合成した音声、 j はその速度で収録した音声を与えた場合である。防音室においては 0 km/h 時ではやや離れている傾向があるがクリーン音声と比べると速度の上昇につれ、距離が離れる度合いが少ない。テストコース走行時には 0 km/h 時には距離が近い。他の速度においても合成音声の方がクリーン音声よりも距離が小さくなっている。しか

し，速度の上昇に伴い距離が離れていっていることがわかる．これはクリーン音声の時と同様，走行雑音の混入による影響と思われる．

6.4 考察

図 6.5 に図 6.1 から図 6.4 の各平均を抜き出したものを示す．同じ収録環境同士で比較すると合成音声はクリーン音声に対し，全体的に値が低くなっている．これは収録音声に対し，特徴量ベクトルの分布距離が離れるのを低下させていると言える．そのため，音響モデルの学習においてクリーン音声をそのまま利用するよりは有用であることが考えられる．テストコース走行時においても合成音声はクリーン音声より値が低い．しかし，防音室に比べて値の上昇が激しいことがわかる．これは走行雑音が影響していると思われる．合成音声はクリーン音声をもとに変換しているため，走行雑音は入っていない．よって，合成音声に走行雑音を加えるか，または，テストコース走行時の発話から走行雑音を除去して比較することにより値の上昇を抑えられる可能性がある．

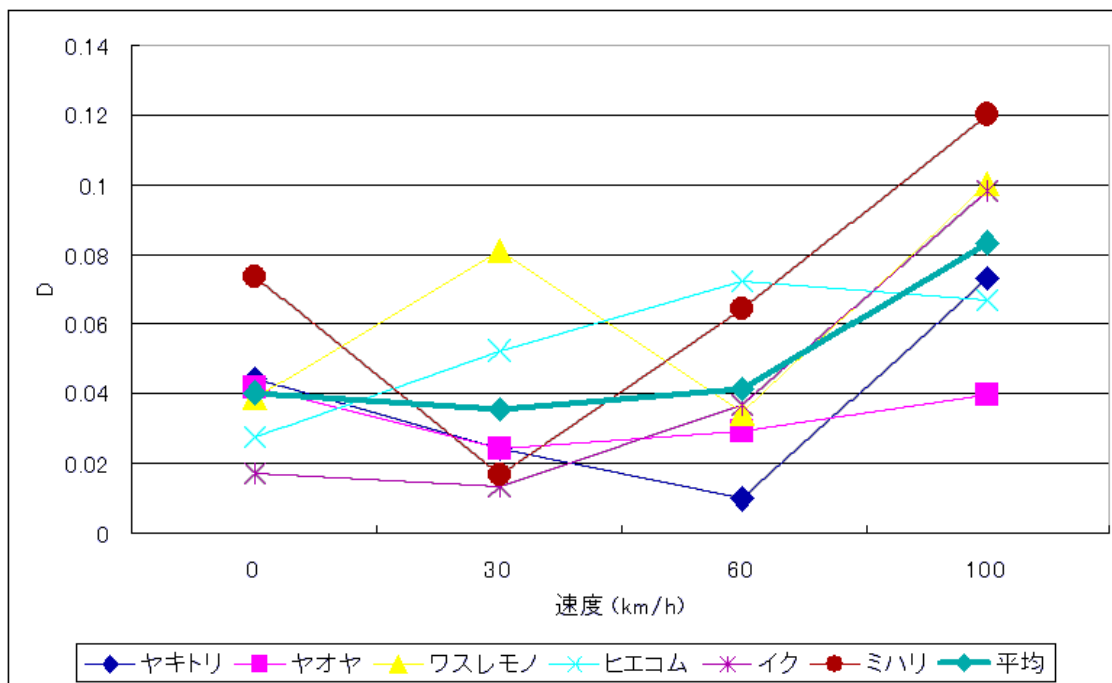


図 6.1: クリーン時と防音室の発話におけるマハラノビス距離

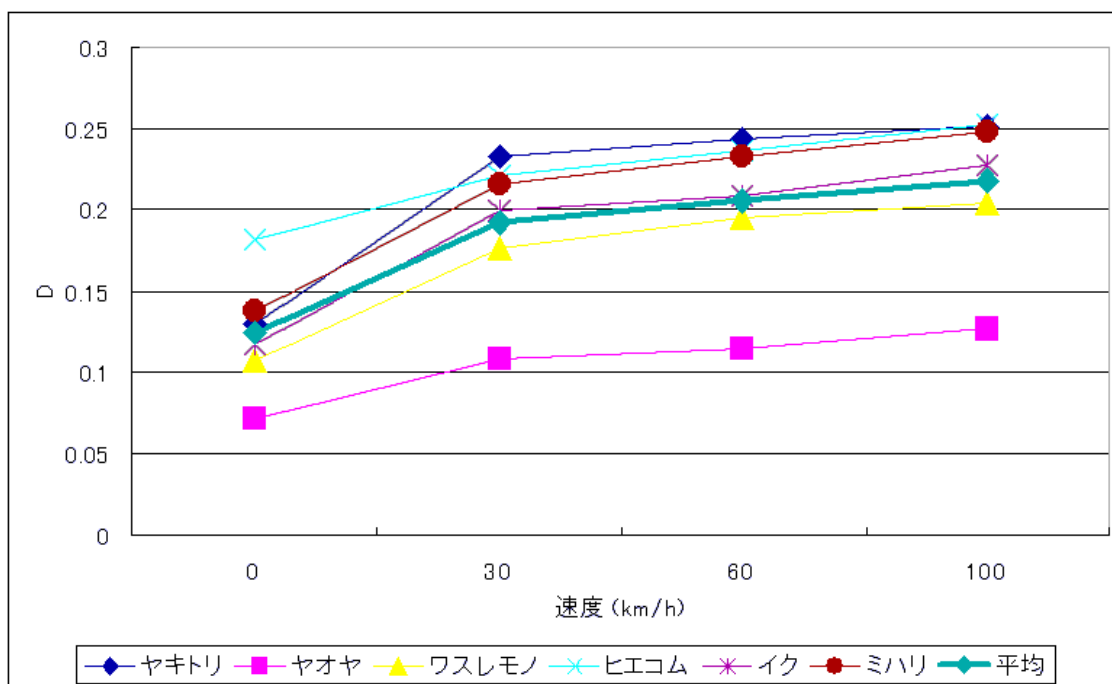


図 6.2: クリーン時とテストコース走行時の発話におけるマハラノビス距離

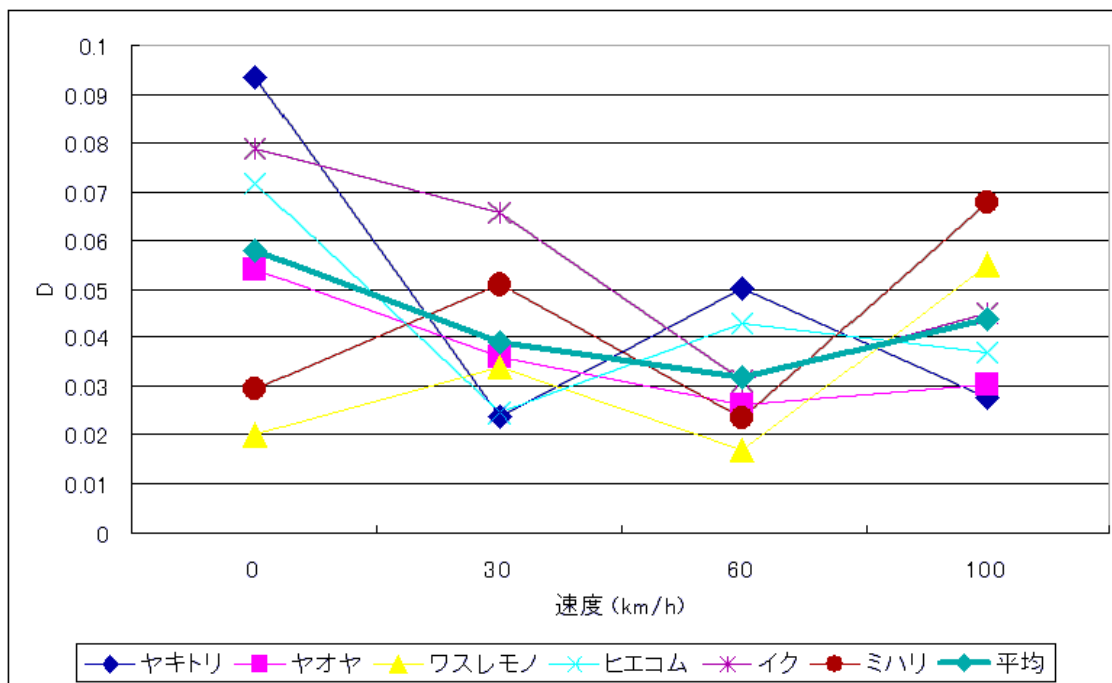


図 6.3: 合成音声と防音室の発話におけるマハラノビス距離

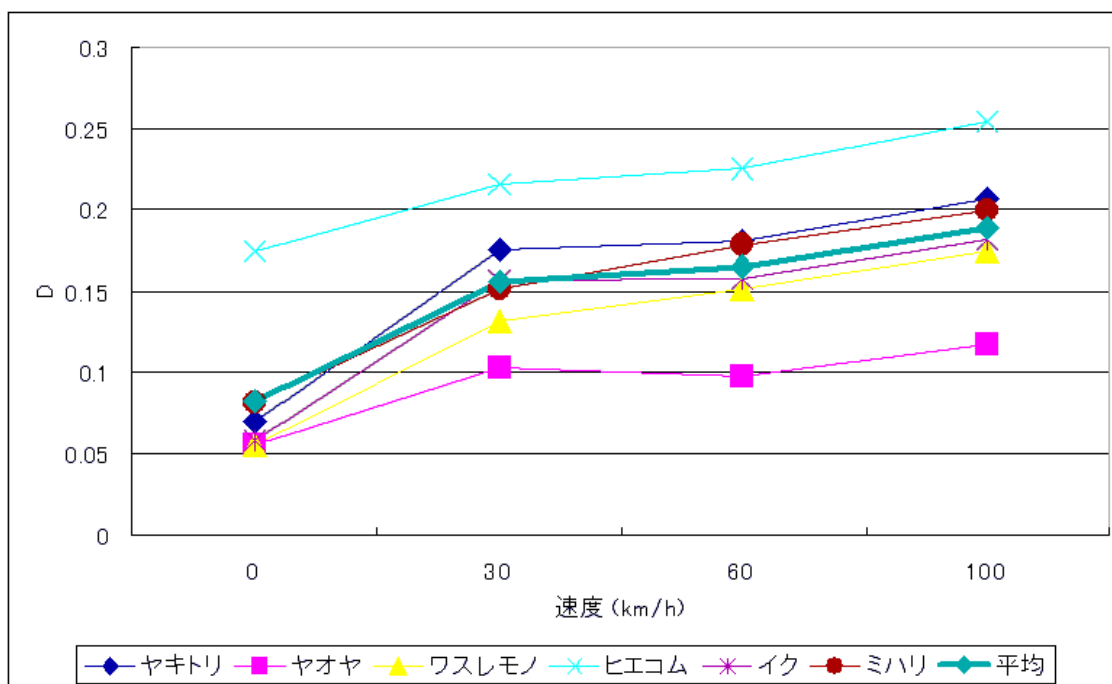


図 6.4: 合成音声とテストコース走行時の発話におけるマハラノビス距離

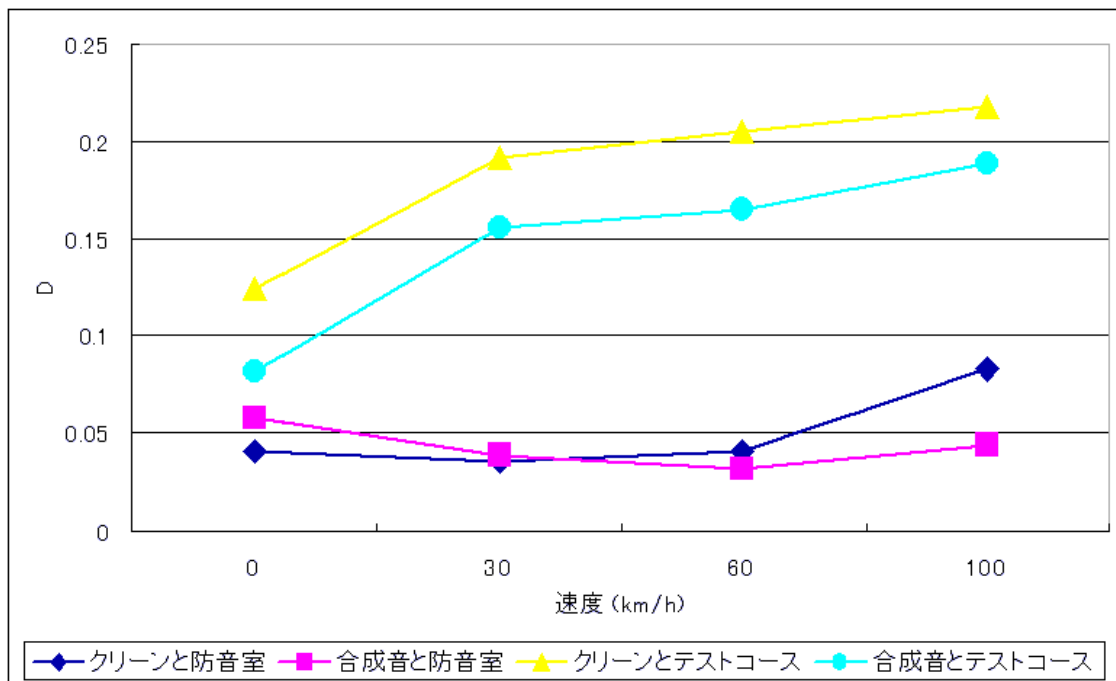


図 6.5: マハラノビス距離の変化の平均

第7章 全体考察

本章では本研究で明らかになった結果について、全体の考察を行う。

- 収録環境の違いについて

今回、実際にテストコースを走行しながらの発話を収録したものと、防音室においてヘッドフォンによって走行雑音を受聴しながらの発話を収録したものの2種類の環境によって収録を行った。Kadambe の報告によると運転しているというタスクを行いながらの発話は音声認識において影響を与えるという [12]。しかし、今回は4章の分析結果に示すようにテストコース走行時と防音室とでは特に変化の傾向に差が見られなかった。これは話者の個人差によると思われる。本研究で収録を行った話者は今回のような実験に慣れている話者が多く、そのために実際に運転しているという緊張感などを感じることなく収録を行えた可能性がある。

- 変換パラメータについて

本研究においては車室内発話音声の収録の際、車種や路面状況などは固定で自動車の速度だけを変化させて収録を行ったことから、速度の値を引数とした音響的特徴量の変換関数を構築した。他に考えられるパラメータとして発話する際の走行雑音の音圧レベルが考えられる。自動車の速度と走行雑音との関連は車種によって異なるため、他の車種への適用を考えた場合、雑音の音圧レベルに対する変換関数を構築するなどの必要があると思われる。

- 音韻の数の違いについて

今回収録した音声では収録の都合上音韻の数に違いがある。そのために変化の傾向に差があらわれた可能性がある。例えば /a/ に関しては数も多く、また、第一モーラとして出現するものが多い。そのために他の音韻に比べて第一フォルマントの上昇が激しいという傾向が出た可能性がある。反対に /e/ に関しては数が少ない。そのためかフォルマント周波数、特に第三フォルマントに対してばらつきが大きくなった可能性がある。

- 評価方法について

今回、評価の方法について音声全体の特徴量ベクトルの分布を測定することによって評価を行った。その結果、合成音声全体としてはクリーン音声に対して音響モデ

ル学習に有用であるという結果が得られた。しかし、第4章の分析結果において音韻によっては値にばらつきがあり、今回作成した変換関数の正当性に疑問が残るものがある。そこで、音声全体にではなく、音韻区間ごとに評価を行うことにより、フォルマント周波数の変換関数など個々の音韻に対して構築した変換関数が正当なものであるか検討することが可能と思われる。

- 音韻のわけかたについて今回、音声区間のうち母音の測定を行い、5母音ごとにわけて特徴量の変化を分析した。しかし、音韻長についてはモーラの順番ごとに分析を行った。これは収録音声を聞いていた際、最終モーラについて強い増加傾向にあることがわかったため、このような分け方をした。そこで他の特徴量についてもモーラごとの順番は関係ないか調べたところ音韻長のような傾向は得られなかった。しかし、発話においてはアクセントの位置において音響的特徴量が鋭く変化することが知られているため、アクセントを考慮した分け方をすると異なる結果が得られた可能性がある。更に、発話において声道の変化が前後の音素によって影響を受ける調音結合と呼ばれる現象が発生するため、同じ母音でも前後の子音との組を考慮した分析を行う、という方法も考えられる。

第8章 結論

8.1 本論文で明らかにされたことの要約

本研究では車室内発話時に生ずる発話歪みの影響を調べるため、音響的特徴量として基本周波数、パワー、フォルマント周波数、スペクトル傾斜、音韻長の変化について自動車の速度ごとに分析を行った。その結果、各特徴量に対し、自動車の速度に応じた変換ルールを作成した。また、変換ルールをもとに車室内発話音声合成システムを構築し、クリーン音声から車室内発話音声を合成した。得られた合成音声は収録音声に対し、クリーン音声よりも音声認識における特徴量ベクトルの分布間距離の差が少ないことがわかった。このことから本研究で構築したシステムを用いてクリーン音声から多数の車室内発話音声を合成することにより、車室内発話音声認識システムの音響モデル学習に役立てられることが見込まれる。

8.2 今後の課題

8.2.1 合成音声を用いた音響モデルの学習

今回は合成した車室内発話音声のMFCCなどの分布を見ることにより音響モデルの学習に有用であるかの評価を行った。しかし、本来は実際に音響モデルの学習を行い、認識率が上がったかどうかをもって評価とするのが望ましい。

8.2.2 子音に関する制御

本研究では音声の母音区間に対して音響的特徴量の変換を行ったが、子音に関してはクリーン時の状態のままである。子音に関しても制御を行うことにより、合成音声を収録音声により近づけることができると思われる。しかし、子音に関してはその種類によって今回測定した音響的特徴量を求めることが困難である。また、子音の種類ごとに変化が激しかったりあまり変化しないなどのことが考えられるため、子音の種類ごとに分析を行うのがよいと思われる。

謝辞

本研究を行うにあたり，終始多大なるご指導・御鞭撻を受け賜りました北陸先端科学技術大学院大学 情報科学研究科 赤木 正人 教授，鵜木 祐史 助教授に心から深く感謝の意を表します．並びに本研究の過程において多くの御協力を賜りました，本学の教官の皆様に感謝いたします．

本研究において用いた車室内発話音声の収録には，日産自動車総合研究所，神沼充伸氏および山崎勝氏の協力を得ました．深謝致します．また，本研究を進める上で有益な議論・助言をして頂いた赤木 研究室の西本 博則 氏，齋藤 毅 氏，羽二生 篤 氏，黄 純芳 氏に厚く御礼申し上げます．そして，研究を円滑に行う助けを頂いた赤木 研究室の皆様に深く感謝致します．最後に，2年間の学生生活を支えてくださった全ての皆様に厚く感謝いたします．

参考文献

- [1] 吉田勝, 小畑秀文, “ ロンバード効果を考慮した低品質単語認識に一手法, ” 音楽講論, 1-Q-21, pp.183-184, Oct.1994.
- [2] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄, “ 音声認識システム, ” オーム社
- [3] 中村哲, 鹿野清宏, “ 認識技術の進展, ” 情報処理学会誌, vol.38, no.11, pp.998-1006, 1995.
- [4] D.B.Pisoni, R.H.Bernacki, H.C.Nusbaum and M.Yuchtman, “ Some acoustic-phonetic correlates of speech produced in noise, ” Proc.of ICASSP, pp.1581-1584, Apr. 1986.
- [5] 鈴木忠, 阿部芳春, 中島邦男, “ 発声変形モデルを用いた騒音環境下音声認識, ” 日本音響学会研究発表会講演論文集, 1-5-1, pp.1-2, (1995-3).
- [6] 鈴木忠, 阿部芳春, 中島邦男, “ 高騒音下における発声変形音声の認識, ” 日本音響学会 講演論文集, 1-7-7, pp.13-14, (1994-3).
- [7] J.H.L.Hansen and O.N.Bria, “ Lombard effect compensation for robust automatic speech recognition in noise, ” Proc.ICSLP, pp.1125-1128, 1990.
- [8] Hideki Kawahara, Ikuyo Masuda-Kasuse and Alain de Cheveigné, “ Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds, ” Speech Communication, 27, pp.187-207, 1999.
- [9] B.J.Stanton, “ Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions, ” Proc. ICASSP, pp.331-334, 1988.
- [10] 中村匡伸, 岩野公司, 古井貞熙, “ マハラノビス距離を用いた日本語話し言葉音声の音響的特徴の分析, ” 日本音響学会 講演論文集, 2-1-4, pp.230-231, (2005-3).
- [11] HTK Speech Recognition Toolkit, <http://htk.eng.cam.ac.uk/>
- [12] S. Kadambe, “ Study Of Effect Of Speaker Variability And Driving Conditions On the Performance of an ASR Engine Inside a Vehicle, ” Biennial on DSP for in-Vehicle and Mobile Systems Sesimbra, Portugal, September 2-3, 2005.

本研究に関する研究業績

口頭発表

竹山佳成, 鵜木祐史, 赤木正人, 神沼充伸, “ 自動車走行雑音下における車室内発話音声の合成, ” 日本音響学会 春季研究発表会, 1-Q-18, March 2006 (発表予定).

Takeyama, Y., Unoki, M.m Akagi, M., Kaminuma, A., “ Synthesis of mimic speech sounds uttered in noisy car environments, ” NCSP, March 2006 (to be appear).