数理解析研究所講究録
第1928巻 2014年 27-57

聴覚におけるスケール分析のための末梢系フィルタバンクのウェーブレット性と非線形性

入野 俊夫 * 河原英紀 * Roy D. Patterson†
* 和歌山大学 システム工学部
† ケンブリッジ大学 生理発達神経科学科 CNBH

概要. 日常生活で、一声聞くだけで, 話者が大人か子供か, おおよその話者の寸法 (身長) がすぐわかる. 同時に話者の寸法に無関係に発話内容 (言語情報) を聞き取ることができる. ごく当たり前のことに見えるが, 話者の寸法が異なると相関が高い声道長も異なり, 音響的な共鳴周波数も異なる. しかしこのスペクトル上の違いの影響を受けないのである. このことから, 人間の聴知覚系には, 寸法 (スケール) と声道形状 (音韻性) を分離抽出する機構があると考えている. この聴覚の計算理論として, 安定化ウェーブレット-メリン変換を提案した. 知覚実験を通してその特性も明らかにしてきた. さらに, この理論を受けて, 通常発声やささやき声の聴覚的スペクトルから声道長推定を行う問題を検討した. この結果, 聴覚末梢系のウェーブレット性と非線形性を反映させたガンマチャープ聴覚フィルタバンクが最も性能が良かった. すなわち, 実際の音声からの声道長スケーリングの推定問題では, 制約付きの非線形性がある方が, 線形のウェーブレット変換よりも良くなるのである. これらの背景と結果を紹介し, 聴覚的非線形性も含めた理論的枠組みの議論の出発点を提供する.

Nonlinearity and Wavelet property of the auditory filterbank suitable for scale analysis in the auditory system

Toshio Irino*, Hideki Kawahara*, and Roy D. Patterson†
*Faculty of Systems Engineering, Wakayama University
†Centre for Neural Basis of Hearing, Department of Physiology, Development, and Neuroscience, Cambridge University

Abstract.

We hear vowels pronounced by adults and children as approximately the same although the vocal tract length (VTL) varies considerably from group to group. At the same time, we can identify the speaker group. This suggests that the auditory system can extract and separate information about the size of the vocal-tract from information about its shape. We had proposed a computational theory, named Stabilized Wavelet-Meliin Transform (SWMT), to explain the observation. Recently, we performed a VTL estimation experiments using the knowledge of the theory. We found that the nonlinear auditory filter bank , which was estimated by psychoacoustical measurement, was better than any other linear filterbanks including wavelet-like one. This implies the problem of the VTL estimation in real speech sounds is not solely the issue of the scale estimation which can be dealt with the wavelet transform. In this paper, we introduce the background and results for the discussion of the theoretical framework including the auditory nonlinearity.

1. はじめに

音声（有声音）は, 音響管である声道を声帯音源によって駆動することによって生成される. これは,「ソースフィルタモデル」と呼ばれる. 母音の違いは, 声道の形状を変えることにより生成される. たとえば,「ア」は舌を後ろにし,「イ」は舌を前に置くことにより発声されていることは, 舌の位置に気をつけると自ら確かめることができる.[*1] 発声された音声の波形をスペクトル分析すると, 複数ある共振のピーク（ホルマント[*2]）の分布が母音ごとに異なる. 一方, 大人でも子供でも, 同じ母音「ア」は,「ア」として発声できるし, 聞き取ることもできる. ところが, 頭の寸法により声道長 (Vocal Tract Length, VTL) が異なるため, ホルマント周波数自体は大人と子供で異なることになる. 初歩の物理学で教えるとおり, 音響管の長短で共鳴周波数が変わり, スペクトルのピーク周波数は, 音響管長に反比例するスケール関係になる. この音声を聞いた聴取者は, 何らかの手段でその違いを正規化し, 同じ「ア」と聞いているはずである. すなわち, 声道長 (Vocal Tract Length, VTL) を正規化し, それぞれの母音に特有な特徴量の分布を揃える処理が, 人間の知覚系に備わっていると考えられる.

この声道長正規化は, 不特定話者の自動音声認識で有効な手法とされ, Wakita [1] 以来, 数多くの研究があり, 様々な方法が提案されている. また, 最近, 声道長正規化による２話者間の音声モーフィング（特に男女間）の音質が改善されることが報告されている [2,3]. これらの基本となっている手法では, 短時間フーリエ変換の直線周波数軸をメル周波数軸等の疑似対数軸に周波数ワーピング関数等により変換し, その上でスペクトルシフトやシフト不変変換を行う. ここで, たとえば, スペクトル表現の選択／改善や, 正規化の係数の推定法、あるいは学習法をどのようにするかが議論の対象となってきた.

これらに対し, 音声から各々の話者の声道長自体を推定する問題には関心が集まっておらず, 研究もそれほど行われていない. しかし, 上記の音声モーフィングでは単に正規化をするだけではなく, 目標となる話者の声道長情報も必要である.

本稿では, 聴覚系における声道長推定／正規化の理論 [4–6] や, それを支持する聴覚心理実験に関して紹介する [8–11]. さらに, この聴覚的な知見を聴覚フィルタバンクレベルで導入した声道長推定手法と有効性について紹介する [3,12–16]. 有声音ばかりではなく、無声音のささやき声においても、同一話者の通常発声の場合と同様に声道長推定を行って比較した. さらに, 音声から推定された声道長と身長との関係を, 磁気共鳴画像 (MRI) から得られている声道長と身長との関係と対比し, 推定の妥当性を検討した. 結果としては, 心理物理実験データを反映する非線形性の入ったガンマチャープ聴覚フィルタが, 線形

[*1] いままでそのようなことに気を配ることもなかったであろうが, ぜひ試していただきたい.

[*2] 音声学での呼び方 formant を日本語でこのように表記する.

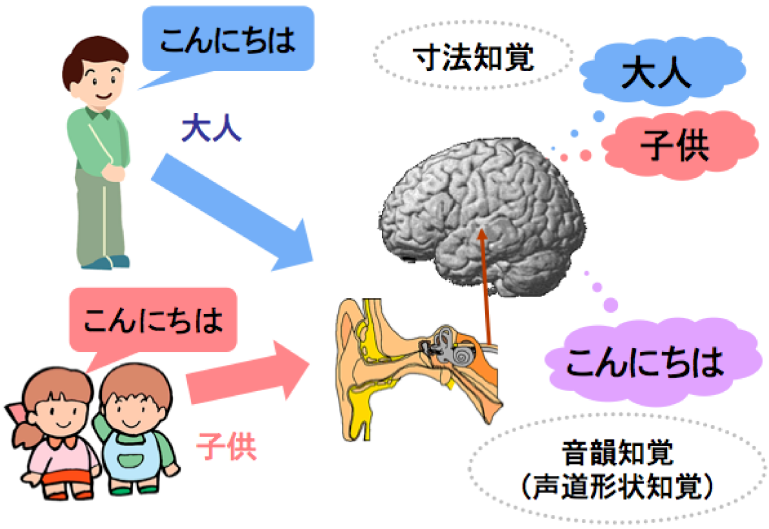

Fig. 1. Size and Shape perception from sound source.

のフィルバンクよりも推定精度が良いことがわかった. ここでは, 聴覚フィルタバンクの非線形性についても述べ [17–21, 23], 線形のウェーブレット的性質から拡張するための理論構築の出発点を提供したい.

2. 聴覚系における寸法-形状知覚と理論

2.1 聴知覚特性

図 1 に示すように, 大人と子供が発声した同じ言葉を聞いたとき, これらの音声はスペクトル分布としては異なっていても同じ言葉として知覚することができる. また同時に話者が大人か子供かを認識することが可能である. このことから, 人間の初期聴覚系において, 聞いた音から発音体の寸法（=声道長）情報と形状（=声道の断面積関数）情報に分離し, 抽出する機能があるという仮説を立て, 理論を提案している [4–6] . これを受けて聴知覚実験も行われ, 寸法の弁別閾はおおよそ 5% 程度であることがわかっている. さらに, 通常発声範囲をはるかに超えた基本周波数-声道長の組み合わせの合成音やささやき声においても, おおよそ 5% の弁別閾は変わらないことや, 音韻や単語の正解率が十分に高いことがわかっている [8–11] .

2.2 聴覚計算理論

上記の知覚特性を説明するために, 初期聴覚系で寸法情報と形状情報の分離抽出を行っているという計算理論を提案している [4–6]. 図 2 に, このアルゴリズムである安定化

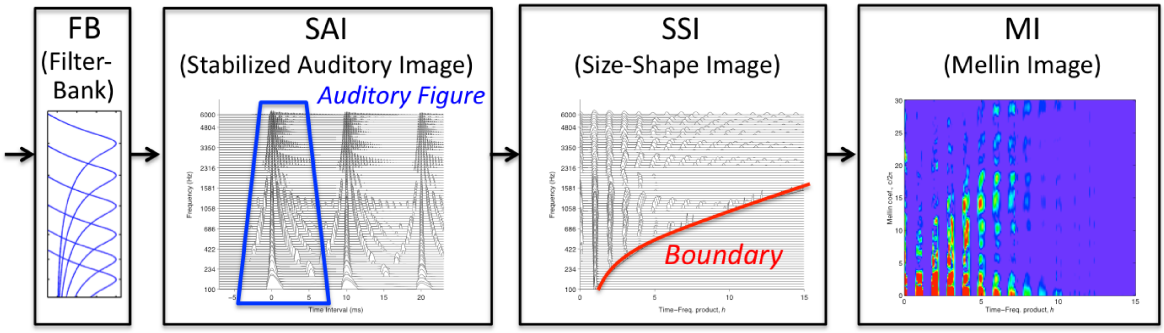

Fig. 2. Computational theory of the size-shape perception: Stabilized Wavelet-Mellin Transform

ウェーブレット-メリン変換のブロック図を示す. 各々のステージについて, 次節以降で説明するが, まとめると以下のようになる.

1. 聴覚フィルタバンク (Gammachirp Filterbank GCFB)
   - 聴覚末梢系で行われる周波数分析を行う.
   - 入力音圧に適応的に増幅度（フィルタ利得）を変える非線形性がある.
   - きわめて小さい音 (0dB SPL) から大きい音 (100dB SPL) まである外界の音を, 聴神経で対応できる 30dB ほどの範囲に納める役割をする.
   - フィルタ利得の入出力関係から「圧縮特性」と呼ばれる.
2. ストローブ時間積分による安定化聴覚像 (Stabilized Auditory Image, SAI)
   - 聴覚系には時間的な積分作用がある.
   - 同時に時間的な微細構造 (Temporal Fine Structure, TFS) も保持される.
   - 時間積分は, 通常漏洩積分器等のスムージングフィルタが説明に用いられるが, TFS は消えてしまう.
   - この相矛盾する条件を同時に満足させる手法がストローブ時間積分である.
   - この処理の結果得られた表現が安定化聴覚像である.
3. スケール共変性表現 (Size-Shape Image, SSI)
   - 1 周期分に相当する境界線以上の部分を聴覚図 (Auditory Filgure, AF) と呼ぶ.
   - 母音の違いにより聴覚図は変わる.
   - 寸法の違いは, 聴覚図の垂直方向の位置の違いとしてだけ表される.
4. スケール不変表現 (Mellin Image, MI)
   - Mellin 変換を取ることにより, スケール変形 (寸法変化) に対して不変な表現が得られる.

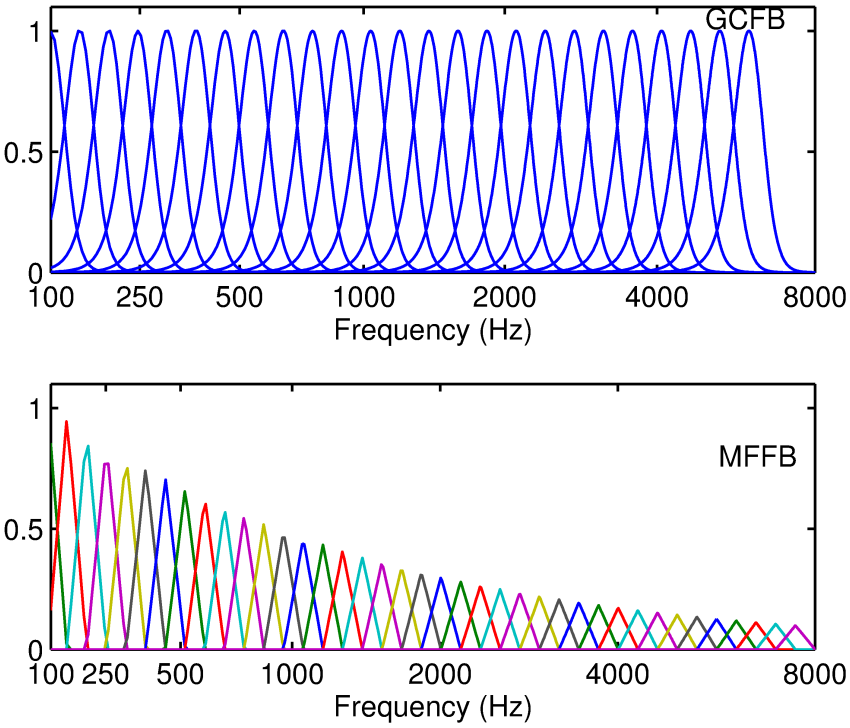

Fig. 3. Characteristics of gammachirp filterbank (upper panel). The number of the filter is restricted for the plot. Characteristics of mel-frequency filterbank (lower panel).

2.2.1 聴覚末梢系のフィルタバンクモデル

入力された音は, 聴覚フィルタバンク (FB) で, 時間軸と $\mathrm{ERB_N}$ 軸 [7](疑似対数周波数軸) を持つスペクトログラム的な分析が行われる. また, 実際に聴神経の活動まで模擬する場合は半波整流を行い, 神経活動パターン (NAP) と呼ぶ表現にする. この聴覚フィルタの周波数特性は, 心理物理実験的に推定できる [17–20]. 推定されたフィルタ特性は非線形を持ち, 入力音圧に依存して周波数特性が変化し, 利得も変化する (圧縮特性を持つ) ことが知られている. これらの非線形性に関しては 4 節で述べるが, 線形の第一次近似としてはウェーブレット変換に似ていると古くから指摘されている [24]. この聴覚末梢系の周波数分析に関しては研究の歴史は長く, 古典的な機械振動解析から, 単純ではあるが見通しの良いフィルタバンクまで, 数多くのモデルが提案されている [25]. フィルタバンクの周波数特性の一例を, 図 3 上図に示す.

音響管の寸法が変化すると, インパルス応答が時間的に伸縮される, スケール変形となる. この音のスケール変形に対して, フィルタ系による歪みを与えないという意味では, 線形のウェーブレット変換が最も良い. これは, どのフィルタも同じインパルス応答 (kernel 関数) でスケールのみが違うため, 外界の音がスケール変形しても必ず同じ形のフィルタ

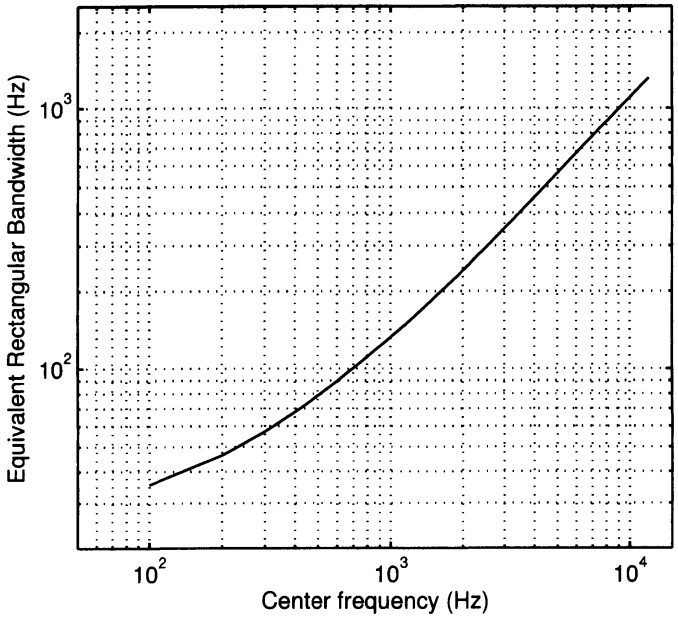

Fig. 4. The relationship between the center frequency and the bandwidth estimated by psychoacoustic experiments. This is used for the gammatone filter. The bandwidth for the gammachirp filter is about 1.5 times.

で処理されるからである. このウェーブレット変換では, 周波数と帯域幅が比例する定 Q 特性が成立していることが必要条件となる. 聴覚心理実験の結果から得られた, フィルタの中心周波数と帯域幅の関係を図 4 に示す. 縦軸は, 聴覚心理物理でよく用いられる等価矩形帯域幅 (Equivalent Rectangular Bandwidth, ERB) で, 健聴者の ERB_N (Hz) はフィルタの中心周波数 f_c (Hz) に対し次式で与えられる [17].

$$ERB_N = 24.7 \cdot (4.37 * f_c/1000 + 1).$$

この図を見ると, おおよそ 500Hz 以上において周波数と帯域幅が比例し, 定 Q 特性を満足していることがわかる. すなわち, その領域ではウェーブレット変換を用いてフィルタ系を構成できることになる.

フィルタバンクを構成する各チェンネルのフィルタ (kernel 関数) としては, ガンマトーン (gammatone) の系統が最も有力である. このガンマトーン[*3]は, 生理実験で得られたネコの基底膜振動のインパルス応答を近似するための実験式として元々提案されたものである [26]. その後, 様々な変遷を経て, 現在まで最も良く使われるフィルタ系となっている. この中には, Lyon が提案した one-zero gammatone や Meddis らの DRNL, Irino and Patterson のガンマチャープ (gammachirp) などがある (経緯や文献は [21–23, 25, 27] 参照).

[*3] ガンマトーン (gamma-tone) は, 包絡線がガンマ関数 (gamma) で, 搬送波が正弦波のトーン (tone) であることからの造語である.

このガンマチャープ*4. は, 以下で述べる初期聴覚系の内部表現（スケール表現）の考察に踏み込み, Mellin 変換 (3.2.4 項参照) と時間 (間隔) 軸で張る空間の最小不確定性を持つ関数として関数解析的に求められたものである [21] . Appendix A にその導出を示す. ガンマチャープの特殊解であるガンマトーンも含めた聴覚フィルタは, 音源の寸法やスケール変形を扱う情報処理に最適な系を構成していると解釈することができる.

2.2.2 初期聴覚系における時間積分と安定化聴覚像

音量の小さな短音の数を増やしていくと, 聞こえる音の大きさ（ラウドネス）が徐々に大きくなることが知られている. これは, 聴覚系に時間積分の機能があることを示している [9]. この説明モデルとして, 時間窓をかける形の積分（スムージング）が従来使われてきた. しかし, 人間は時々刻々変化する微妙な音色も, 同時に聞き分けることもできる. そこで, この時間的な詳細構造 (temporal fine structure) を保持する機構が別途必要となってしまう. 音の大きさ知覚や微細構造知覚といった現象ごとに別個の説明モデルを作ることは, 複雑になるだけで本質から遠ざかる可能性が大きい.

そこで時間積分の特性を持ちつつ時間的詳細特性の保持するために考えだされたのが, ストローブ時間積分 (Strobed Temporal Integration, STI) である [5, 7, 25]. これは, 振動体をストロボスコープ*5を用いて撮影した場合や, オシロスコープの同期モードで波形を見る状況と類似のものと思えば良い. 聴覚モデルにおいては, 各々の周波数チャンネルごとに, ある時点の神経活動パターンを, 時間間隔と周波数の軸を持つ 2 次元のイメージバッファにピーク時点を同期させながら積分する. たとえば, 音声であれば基本周期ごとに類似な神経活動パターンが繰り返される. これをピッチパルスに同期して積分する. これで得られる表現を, 安定化聴覚イメージ (Stabilized Auditory Image, SAI) と呼ぶ（図 2 の 2 ブロック目). この 2 次元イメージは, 入力が音声ならば定常母音では定常的で, 音節の移り変わりで変化する動画的な表現となる. この SAI の上では, 基本周波数ごとに同じ活性度パターンが繰り返される. この 1 周期分が, 話者の声道の共振特性を示す聴覚図 (Auditory Figure, AF) である.

この聴覚図 (AF) を用いれば, 話者の発声している音韻や話者の声道長を安定に推定できるはずである. 一方, 音声に留まらず, この SAI の 2 次元表現上での様々な場所で特徴ベクトルを取り, Web 上の音検索に使う試みも最近提案されている [28].

---

*4 ガンマチャープ (gamma-chirp) は, 包絡線がガンマ関数 (gamma) で, 搬送波が周波数変化のあるチャープ波 (chirp) であることから命名された [21] このガンマチャープ関数は Gabor 関数同様, 初期位相を適切に選ばない限り周波数 0 で値が 0 にならず admissible 条件を満たさないため, 厳密な意味でのウェーブレットカーネルとはならない. 聴覚系自体に合成系は無いので, 条件を緩めた「半ウェーブレット」的な扱いがあれば良いのかもしれない. もっとも, 音声処理に関しては低い周波数 (50 Hz 以下) は無視できるので, ガンマチャープでも実質的に分析合成系を構成できる.

*5 ストロボ/光源を一定間隔で一瞬発光させる装置. 振動体の振動周期に同期させると静止画撮影も可能である.

2.2.3 スケール共変性表現

この聴覚図 (AF) は, 外界の音がウェーブレットフィルタに畳み込まれて出てきた表現を安定化させた信号表現である. ここで, この聴覚図 (AF) を, 縦軸のチャンネルごとに中心周波数に逆比例させて時間間隔軸を伸縮することを考える. 各ウェーブレットフィルタはこの伸縮によりインパルス応答が同一の kernel 関数に正規化される. フィルタ自体はすべて同一となるので影響は無視出来て, 伸縮された聴覚図は外界の音の特徴をそのまま表現していることになる. ここで 得られた表現を寸法形状イメージ (Size-Shape Image, SSI) と呼ぶ (図 2 の 3 ブロック目). この表現上では, 音声における声道長伸縮 (スケール変化) の効果は, 伸縮の無い同一パターンの上下移動として単純化されて表現されることになる. これがスケール共変表現である.

ここで, 低い周波数側では, SAI において聴覚図 (AF) が基本周期ごとに重なることに注意が必要である. SSI を取るときにこの重なり部分が切り捨てられるため, 図 2 の 3 ブロック目に示した, 境界線 (Boundary) の下側に活性度が無い空白部分ができる. 左端のストローブしたピッチパルス時点から離れるに従い, 有効なパターンの下限周波数が高くなる. この空白部分は, 本来音源がインパルスであれば表すことのできた声道特性が, 声帯振動の基本周期 (基本周波数 F_0 の逆数) の影響により表現できない所である. これは, 声道の音響管を短い周期の声帯振動によって駆動する音声生成過程の避けがたい特徴である. 声道長を安定に推定するためには, 音響管の共振特性と駆動源の励振特性をスペクトル情報から上手に切り分ける必要がある.

2.2.4 スケール不変特徴

最終段は, SSI の縦方向にフーリエ変換をし絶対値を取って寸法を正規化した, メリンイメージ (Mellin Image, MI) である (図 2 の最終ブロック). フィルタバンクの対数軸上でフーリエ変換を行なうことはメリン変換に相当する. この時, 寸法情報は位相項として得られる. この処理は, 大脳の一次聴覚野で表現されている周波数軸に順序よく並んだトノトピー表現空間から, 周波数成分を取り除き, さらに内部の処理に進む段階であると想定している. この意味で, このメリンイメージは, Shamma の提案する大脳皮質の受容野 (Receptive Field, RF) [29, 30] の一部を表現していると位置づけられるかもしれない. 逆に言えば, RF の処理の中には Mellin 変換として定式化できるものがあるものと考えられる.

3. 聴覚フィルタバンクによる音声からの声道長推定

話者の寸法 (声道長) を安定に推定するためには, 2.2.3 節の初期聴覚系理論の聴覚図 (AF) の考え方を取り入れることが重要となる. しかし, この聴覚末梢系より内部の処理に関しては, 生理学的な観測データが無いため議論の余地がまだ残っている. これに対し, 聴覚末梢系を近似するフィルタバンクに関しては, 多くの知見に基づき近似の度合いに応じ

て多種提案されている. そこで, どのようなフィルタバンクや分析条件が, 声道長推定に最も有効かという問題に置き換えて考える. これにより, 効率の良い末梢系表現を考察することができる.

3.1 フィルタバンクと周波数領域の選択

「聴覚的」と称するフィルタバンクは数多く提案されている. この中から, 最も良いものを選ぶ必要がある. さらに, 前節で述べたように低い周波数領域は基本周波数や音声の駆動音源波形の形状の影響を受ける. 特に聴覚図 (AF) で表現できる下限周波数には注意が必要である. また, 逆に高域は個人性の影響が大きく, 例えば 4〜5kHz 付近に梨状窩による零が存在する場合もある [31]. この 2 つの領域に挟まれた間に, 声道長情報が最も良く表われる領域があるはずである. そこで周波数帯域の選択によって, 推定誤差がどのように変化するかを調べ, 誤差最小となる条件を設定する必要がある.

3.2 声道長推定手法

声道長推定手法の詳細は別報告 [3, 12–16] に譲り, 本節と付録 Appendix B で概要を述べる.

3.2.1 2 話者間の声道長比の推定

2 人の話者 i, j を設定する. 一般に声道長が異なるためスペクトル分布が異なる. そこで, 片一方のスペクトル S_j をスケール伸縮の r 倍をし, もう片一方の話者のスペクトル S_i と最もマッチングする所を探すことを考える. そこで, 2 つのスペクトルの距離が最小となるスケール伸縮比率 $r_{i,j}$ を, その 2 人の話者 i, j の組み合わせにおける声道長比の推定値とする.

3.2.2 全声道長比の推定

男女計 N 名の話者間の声道長の比を, 総当たりで推定する. 全組み合わせだけでなく, 処理の順番も考えたため, ${}_NP_{N-1}$ 通りとなる. 通常発話のみを使った実験では $N = 28$ で, 順列は ${}_{28}P_{27} = 756$ 通りである. 比較対象の 11 種類のフィルタバンク, 計 56 種類の周波数帯域について, 3 文章を用いて実験を行った. ささやき声と通常発話の両方を扱う実験では $N = 21$ で, 順列は ${}_{21}P_{20} = 420$ 通りである. 通常発話のみで最も良かった 2 種類のフィルタバンクを, 計 56 種類の周波数帯域について, 2 文章を用いて実験を行った. スケール伸縮比率 $r_{i,j}$ を求めるアルゴリズムは, 最小化したいスペクトル距離を $D_{spec}(i, j, r)$ として, 以下のように表される. (通常発話のみを使った実験の例)

for $N_{filterbank} = 1 \rightarrow 11$ (for all filtebanks) **do**
  for $N_{sentence} = 1 \rightarrow 3$ (for all sentences) **do**

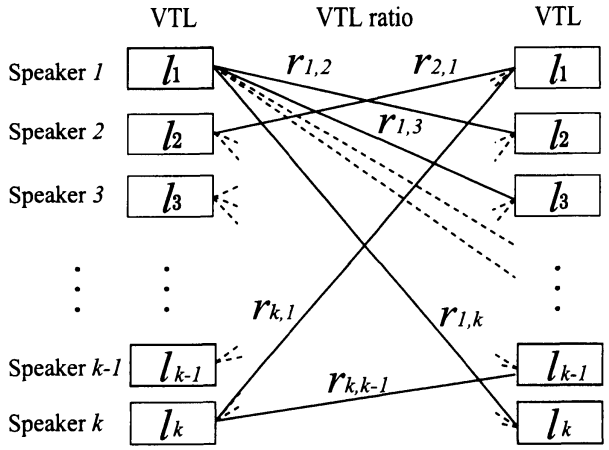

Fig. 5. *VTL ratio $r_{i,j}$ is calculated as a ratio between VTL of i-th speaker l_i and VTL of j-th speaker l_j. All permutation were calculated.*

for $N_{Fregion} = 1 \rightarrow 56$ (for all combinations of frequency region) **do**
  for $i = 1 \rightarrow 28$ (for all speakers) **do**
    for $j = 1 \rightarrow 28, i \neq j$ (for all speakers except for the same) **do**
      $r_{i,j}(N_{filterbank}, N_{sentence}, N_{Fregion}) = \arg\min_{r}(D_{spec}(i, j, r))$
    end for
  end for
end for
end for
end for

声道長比推定を全 140 万回 (=11 × 3 × 28 × 27 × 56) 行う比較的大規模な実験である. この各々の要素について以下で述べる.

3.2.3 最小 2 乗近似

各フィルタバンク・文・周波数領域について, 声道長比 $r_{i,j}$ が求まった時点で, Appendix B に示した手法で最小 2 乗近似を行う [3, 12–16]. 最小 2 乗近似の結果求まった声道長比 $\hat{r}_{i,j}$ と, 元の $r_{i,j}$ の差の rms 値を推定誤差とした. これは, 1 人の話者が 1 つの声道長の真値を持っているとし, 選んだ 2 話者の比を取った値に対して, どの程度ずれるかを測っていることになる. どの話者の組み合わせや, どの発話内容であったとしても, ばらつきが小さければ安定な推定と言うことができる.

理想的には声道長の真値がわかれば良い. MRI 装置を用いた声道断面測定を行えばある程度であるが, 同じ被験者の音声データが必要である. また, 実際の声道長と音声スペク

Table 1. *Filterbanks compared in VTL estimation. See results in Fig. 7.*

Filterbank	Description	#Channel
$GCFB_{dyn}$	dynamic compressive GCFB	100
$GCFB_{lin}$	linear GCFB	100
$GTFB_{025}$	gammatone filterbank	24
$GTFB_{050}$	(linear)	50
$GTFB_{100}$		100
$MFFB_{STR24}$	mel-frequency filterbank	24
$MFFB_{STR40}$	based on TANDEM-STRAIGHT	40
$MFFB_{STR120}$	spectrogram (linear)	120
$MFFB_{STFT24}$	mel-frequency filterbank	24
$MFFB_{STFT40}$	based on STFT spectrogram	40
$MFFB_{STFT120}$	(linear)	120

トルとの関係は, 第 1 次近似としてはスケール関係 (比例関係) が成立するが, まだ詳細には解明されていない. さらに, ここではスペクトルマッチングだけが目的のため, 単純なスケール関係を考えている (Appendix B 参照).

3.2.4 比較対象のフィルタバンク

「聴覚的」フィルタバンクは様々提案されているが, フィルタバンクの種類によりスペクトル表現が異なるため, 性能が異なるはずである. ここでは, ガンマチャープフィルタバンク (GCFB), 広く用いられているガンマトーンフィルタバンク (GTFB), 音声認識で最も用いられているメル周波数フィルタバンク (MFFB) を比較対象 [32] として, 通常発話のみを使った実験では, 以下の 11 条件を設定した. STRAIGHT 以外は, 25ms の hamming 窓でパワーを平均化したスペクトログラムを用いた. ささやき声と通常発話の両方を対象とした実験ではこの内, 最も良い $GCFB_{dyn}$ と $MFFB_{STR40}$ を用いた.

$MFFB_*$ は, 短時間フーリエ変換や STRAIGHT で時間-周波数表現にした上での重み関数である. その意味ではインパルス応答は定義されていない. 図 3 下図に示すように, 重み関数はメル周波数上の三角窓で, これを全て加算すると値 1 の平坦な周波数特性となる. 形式上コンプリートフィルタバンクの形である. これに対し, $GCFB_*$(図 3 上図) や $GTFB_*$ はフィルタどうしの重なりが大きくオーバーコンプリートフィルタバンクの形式になっている. また, $GCFB_{lin}$ は, $GTFB_{100}$ の約 1.5 倍の帯域幅を持つフィルタから構成されているため, オーバーコンプリートネスもさらに高い.

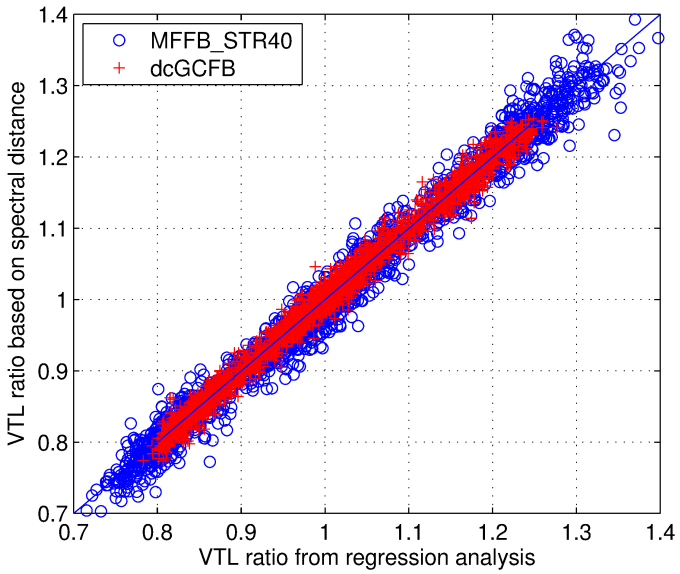

Fig. 6. Relationship between VTL ratios r and $\hat{r}$ estimated using $\mathrm{GCFB}_{\mathrm{dyn}}$ (+) and $\mathrm{MFFB}_{\mathrm{STR40}}$ (○) with best frequency regions.

3.2.5 推定のための周波数領域と評価用音声

聴覚図 (AF) や付録 B.2.1 の知見から, 声道長の推定に用いる周波数領域を制限した方が良い可能性がある. そこでここでは, 様々な周波数領域を検討するため, 下限周波数 100〜800Hz で 100Hz 刻み, 上限周波数 2000〜8000Hz で 1000Hz 刻みで設定した. これらの組み合わせは 8 × 7(=56 通り) のメッシュ状となる. この各点ごとに推定誤差を計算した.

また, 音声サンプルによって, 推定される声道長が異なる可能性もある. そこで, 推定の安定性を評価するために, 複数の話者で, 長さの異なる複数の文章を用いた. 通常発話のみを使った実験では, 話者 28 名 (男女各 14 名) の 3 文（各々 10, 14, 20 音節で構成されている）を用いた. ささやき声と通常発話の両方で評価する実験では, 話者 21 名 (男 14 名女 7 名) の 2 文（各々 10, 14 音節) を用いた. 声道長比は同一の文章を発話した音声サンプル間で計算した.

3.3 通常発話音声を使った実験の結果

図 6 に, $\mathrm{GCFB}_{\mathrm{dyn}}$ (dcGCFB) (+) と $\mathrm{MFFB}_{\mathrm{STR40}}$ (○) で推定した声道長比を示す. 横軸は, 最小 2 乗近似の結果の声道長比 ($\hat{r}$), 縦軸は, 元のスペクトル距離から求めた声道長比 (r) である. また, フィルターバンクごとに最も良く推定された周波数領域での結果を示している. この図から, $\mathrm{GCFB}_{\mathrm{dyn}}$ の方が推定値のばらつきが小さいことがわかる. その分安定に推定できていると考えられる. また, $\mathrm{MFFB}_{\mathrm{STR40}}$ では, 声道長比が 1.3 以上とやや大きい場

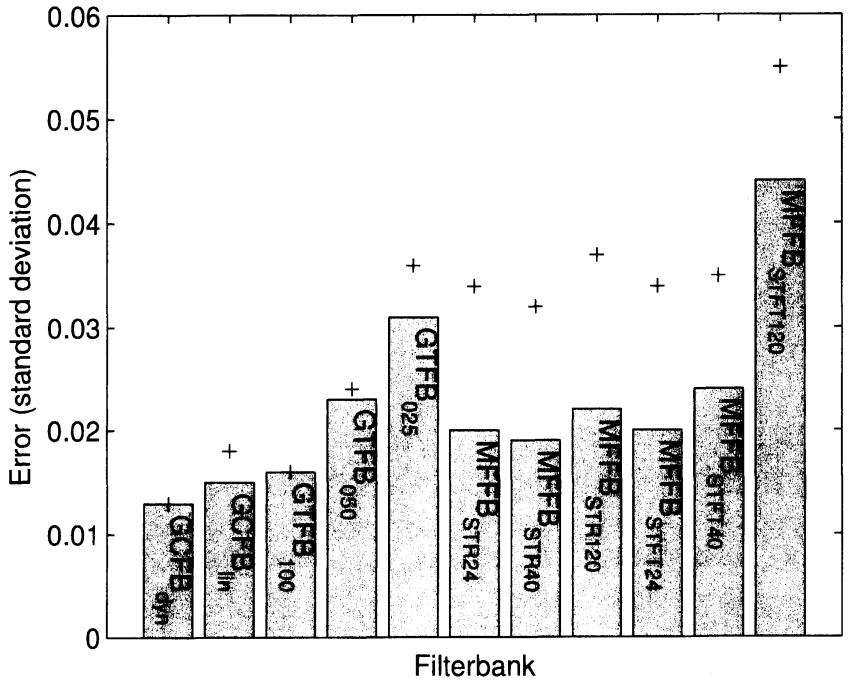

Fig. 7. Estimation error (standard deviation) for the filterbanks. Bar shows the minimum error when the frequency range is properly selected. + shows the error when the frequency region is [500,5000].

合が多い. これに対して, $GCFB_{dyn}$ では, コンパクトな範囲に収まっていて, 推定の精度は高いと考えられる. この妥当性は, 3.4.6 節でさらに検討する.

図 7 に, フィルタバンクの種類ごとに最良周波数帯域を選択した場合の誤差を棒グラフで示す. 周波数領域は, フィルタバンクごとに異なる. この図から以下のことがわかる.

- $GCFB_{dyn}$ の場合最小誤差で, 線形の $GCFB_{lin}$ よりも良い.
- $GTFB_{100}$ は, $GCFB_{lin}$ と同程度である.
- $GTFB_*$ の帯域幅が狭まるにつれ, 誤差は大きくなる.
- $MFFB_{STR24}$〜$MFFB_{STFT40}$ は同程度の誤差で, $GTFB_{100}$ と $GTFB_{050}$ の中間的な値となる.
- $MFFB_{STFT120}$ は, F_0 非依存の STRAIGHT スペクトルを基にした $MFFB_{STR120}$ より格段に誤差が大きい.

表 2 に最小誤差とそれを与える周波数領域を示す.

- どの場合でも周波数領域の下限周波数は 500Hz 以上である.
- $GCFB_{dyn}$, $GTFB_*$ では, 上限周波数が 5000Hz で比較的広い領域となっている.
- $MFFB_*$ では, 周波数領域の上限周波数が, 2000Hz〜3000Hz で比較的低い.

Table 2. Frequency region for minimum error

Filterbank	Freq.Region	Error	Filterbank	Freq.Region	Error
$GCFB_{dyn}$	[700,5000]	0.013	$MFFB_{STR24}$	[500,2000]	0.020
$GCFB_{lin}$	[500,3000]	0.015	$MFFB_{STR40}$	[600,2000]	0.020
$GTFB_{100}$	[600,5000]	0.017	$MFFB_{STR120}$	[600,2000]	0.023
$GTFB_{050}$	[800,5000]	0.028	$MFFB_{STFT24}$	[600,2000]	0.020
$GTFB_{025}$	[800,5000]	0.033	$MFFB_{STFT40}$	[800,3000]	0.026
			$MFFB_{STFT120}$	[800,3000]	0.045

一方, 音声のホルマント情報は, 2000Hz 以上にも存在する（たとえば, 母音/i/や/e/の第 2 ホルマント）. この情報を用いる方が, どのような音環境でも声道長を安定に推定できると, 一般的には考えられる.

このことを検討するために図 7 の＋マークに, 周波数領域を [500,5000] とした場合の誤差を示した.

- $GCFB_*$ $GTFB_*$ では, 誤差は最小値に近い. 最小値を与える領域が近くためでもある.
- $MFFB_*$ では, 誤差は最小値よりも数割以上大きい. すなわち, 2000Hz 以上の領域の情報は, 有効利用できると言うより, むしろ阻害要因となっていることがわかる.

3.4 通常発話音声とささやき声の両方を使った実験結果

ここでは、有声音を含む通常発話の音声とささやき声の音声のそれぞれから声道長を推定した. 同一話者では, 発声法の違いにかかわらず両方の推定値は強い相関があるはずである. さらに, 話者の身長 (寸法) とも相関が高いはずである. そこで, MRI を用いて計測された声道長と身長との関係 [33] とも比較することを行った.

3.4.1 音声データベース

通常発話のみを使った実験で用いた音声データベースは, 元々基本周波数抽出アルゴリズム検討用で通常発話のデータしか登録されていない [34]. そこで, 同一話者が通常発話した場合とささやいた場合の両方の音声を, 防音室で新規に録音した. 同時に話者の身長も記録した. 話者は 21 から 24 才の男性 14 名・女性 7 名の全 21 名である. 身長は, 147.0cm から 186.0 cm であった. 各話者は日本語文, 30 文を通常発話とささやきで発声した。音声は B&K 4003 マイクと Edirol R4-Pro recorder を用いて, モノラル, サンプリング周波数 48 kHz, 量子化 16 bit で収録された. 話者の口元からマイクが約 30cm 離れるように設置

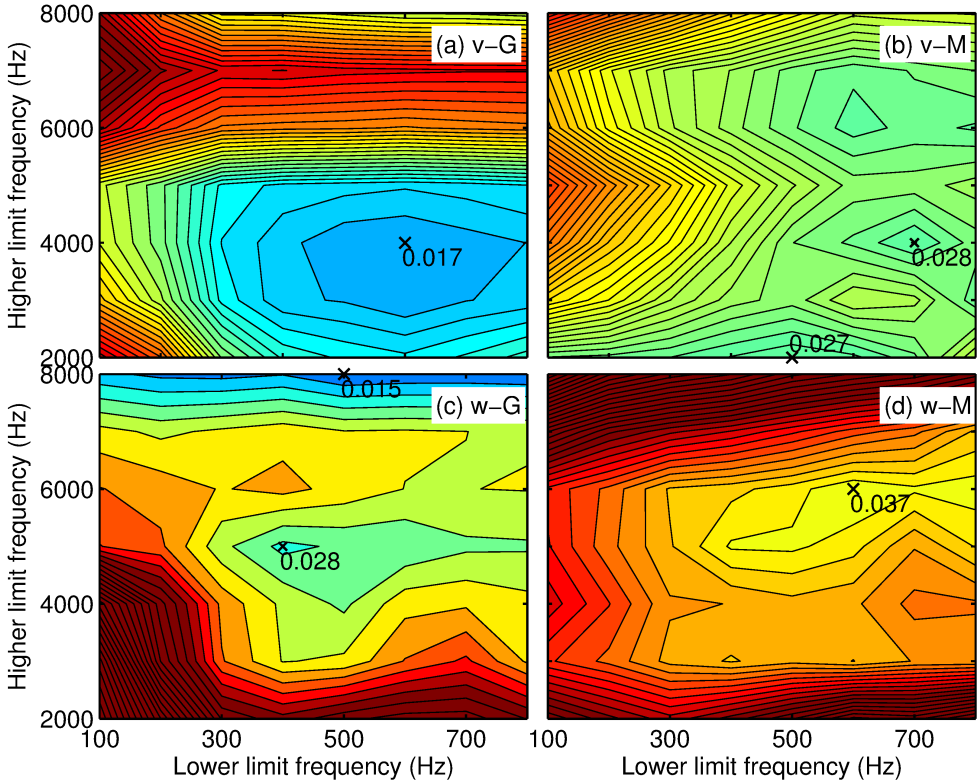

Fig. 8. *Rms error as a function of lower and upper limits of the frequency region* $[f_L, f_H]$. *(a) Voiced speech with* $\mathrm{GCFB_{dyn}}$, *(b) voiced speech with* $\mathrm{MFFB_{STR40}}$, *(c) whispered speech with* $\mathrm{GCFB_{dyn}}$, *and (d) whispered speech with* $\mathrm{MFFB_{STR40}}$. ×: *Global and local minima with error value.*

した.

3.4.2 声道長推定の条件

声道長推定の手法は, 通常発話音声の場合と同じである. 話者数が 21 名のため, 420 ($=_{21}P_{20}$) の声道長比を計算することになる. 以下の実験では, この音声データのうち通常発声の場合と共通の 10 音節と 14 音節の 2 文を用いた. ここでは, 図 7 の有声音の結果から, もっとも誤差の小さかった $\mathrm{GCFB_{dyn}}$ と, 従来から最も用いられている MFFB のうちで最も誤差の小さかった $\mathrm{MFFB_{STR40}}$ を比較することとした.

3.4.3 周波数領域依存性

図 8 に推定誤差の等高線図を示す. フィルタバンクの種類 ($\mathrm{GCFB_{dyn}}$ か $\mathrm{MFFB_{STR40}}$) と発声法 (通常発話かささやき声) の組み合わせで各パネルを表示している. 各々のパネルで, 横軸は選択した周波数領域 (式 Appendix B.1) の下限周波数 f_L, 縦軸は上限周波数 f_H である.

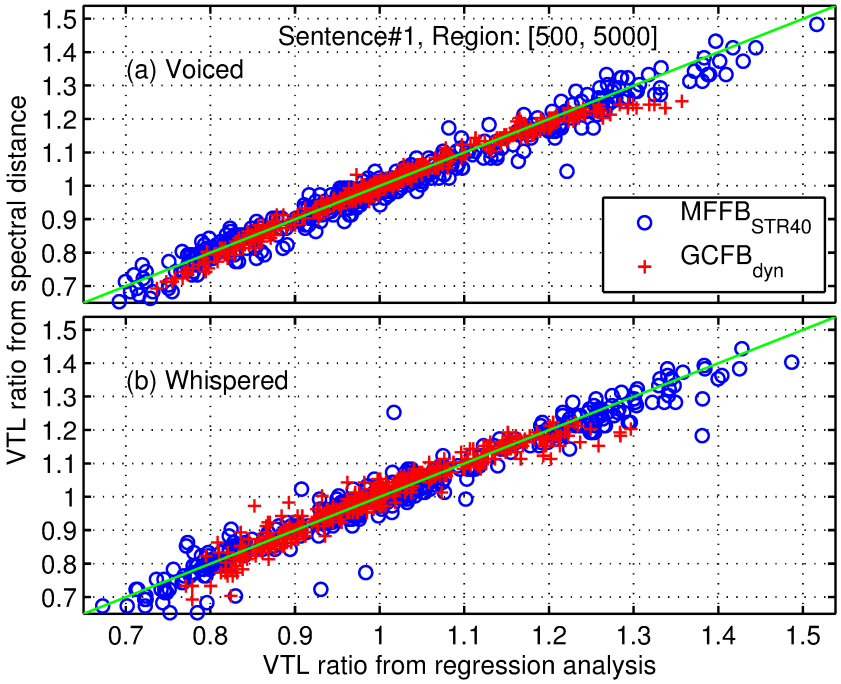

Fig. 9. *Scatter plot of VTL ratios estimated from (a) voiced and (b) whispered speech by using* GCFB$_{dyn}$ *(+) and* MFFB$_{STR40}$ *(o).*

誤差の分布から見ると, 左列の GFFB$_{dyn}$ の方が右列の MFFB$_{STR40}$ よりも全体的に値が小さいことがわかる. 通常発声 (有声音) の場合の最小誤差は, GFFB$_{dyn}$ (図 8(a)) で 0.017, MFFB$_{STR40}$ (図 8(b)) で 0.028 であった. ささやき声の場合の最小誤差は, GFFB$_{dyn}$ (図 8(b)) で 0.028, MFFB$_{STR40}$ (図 8(c)) で 0.037 であった. 発声法にかかわらず, GFFB$_{dyn}$ の方が MFFB$_{STR40}$ よりも推定誤差を小さくできることがわかる. 図 8 の各パネルの軸上にある最小値は計算上の精度の問題があると考えられるので, 中央付近の極小値を与えるものが最も良い周波数領域と考えられる. この最良周波数領域 $[f_L, f_H]$ は,(a) [600, 4000] Hz, (b) [700, 4000] Hz, (c) [400, 5000] Hz, and (d) [600, 6000] Hz であった. したがって, 正確な声道長推定のためには, 周波数領域の下限をおおよそ 500 Hz 以上にすることが良いと考えられる. これは, 有声音の場合には, 声帯振動の基本周波数の高調波成分が, 聴覚スペクトル上で分離されない”unresolved haramonics”になる周波数範囲でもある. 従来研究においては, このような周波数範囲の制限は考慮されていなかったが, この有効性を明確に示すことができた.

3.4.4 声道長比間の関係

図 9 に, 最小自乗分析により得られた声道長比推定値 $\hat{r}$ とスペクトル距離から求めた声道長比 r の間の散布図を示す. この図以降, 分析の周波数領域を [500, 5000] の場合について述べる. この領域では分析フィルタや発声法にかかわらず小さい誤差になる. もし,r が

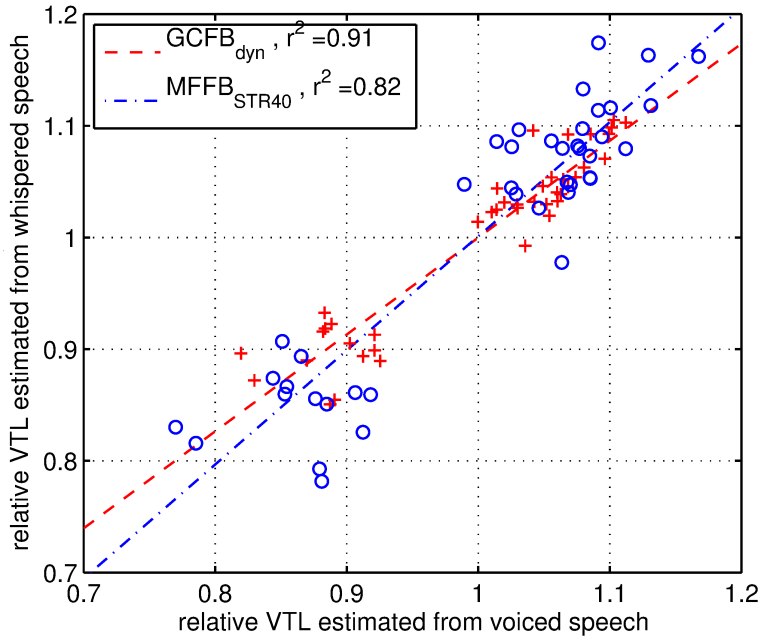

Fig. 10. *Correlation between relative VTLs estimated from voiced and whispered speech. Each point represents VTL combination for one speaker.*

正確に求まっていれば、その最小自乗近似値 $\hat{r}$ と一致し, 対角線上にすべての点乗るはずである. 対角線からのずれが推定誤差に相当する. 図からわかるように GFFB$_{dyn}$(赤 +) の場合が, (a) の通常発声でも (b) のささやき声でも対角線により集中していることがわかる. また, MFFB$_{STR40}$ (青丸) の場合, 大きい声道長比 1.3 以上の推定点が数多くあり, やや信頼がおけない. これは, 3.4.6 節で述べるように, 声道長と身長との間には線形で近似できる相関があり, 最話者の身長比の最大値 1.26(=186cm/147cm) よりもこの値が大きいためである. これに対し, GFFB$_{dyn}$ では, このような外れ値は少なくなっている.

3.4.5 声道長比推定の頑健性

同じ話者が普通に有声発声した場合でもささやき声を出した場合でも, 声道長はほとんど変わらないはずである. 発声法の違いによる共鳴の仕方や音源の違いにより, スペクトル上では若干違いが出てくる可能性はあるが, 少なくとも同一話者ではばらつきは大きくならないと考えられる. すなわち, どちらの発声法の音声を用いても推定される声道長はほぼ同じか高い相関が出ることが予測される. そこで 21 名の話者の通常発声から求めた声道長 $[\hat{l}_1, \hat{l}_2, ..., \hat{l}_{21}]$ とささやき声から求めた声道長との関係を調べた. 図 10 に結果の散布図を示す. GCFB$_{dyn}$ の場合は, 決定係数 $r^2 = 0.91$ で,MFFB$_{STR40}$ の $r^2 = 0.82$ よりも高い相関があることがわかる. このことは, GFFB$_{dyn}$ を用いると発声法によらず, より頑健に声道長を推定できることを示唆している. GFFB$_{dyn}$ の回帰直線の傾きが 1:1 よりもややゆるやかである. サンプル数が 21 人分と小さいためどの程度の精度があるかわからない. しかし, 上記のように発声法の違いによるスペクトル上の違いが反映された可能性も考えら

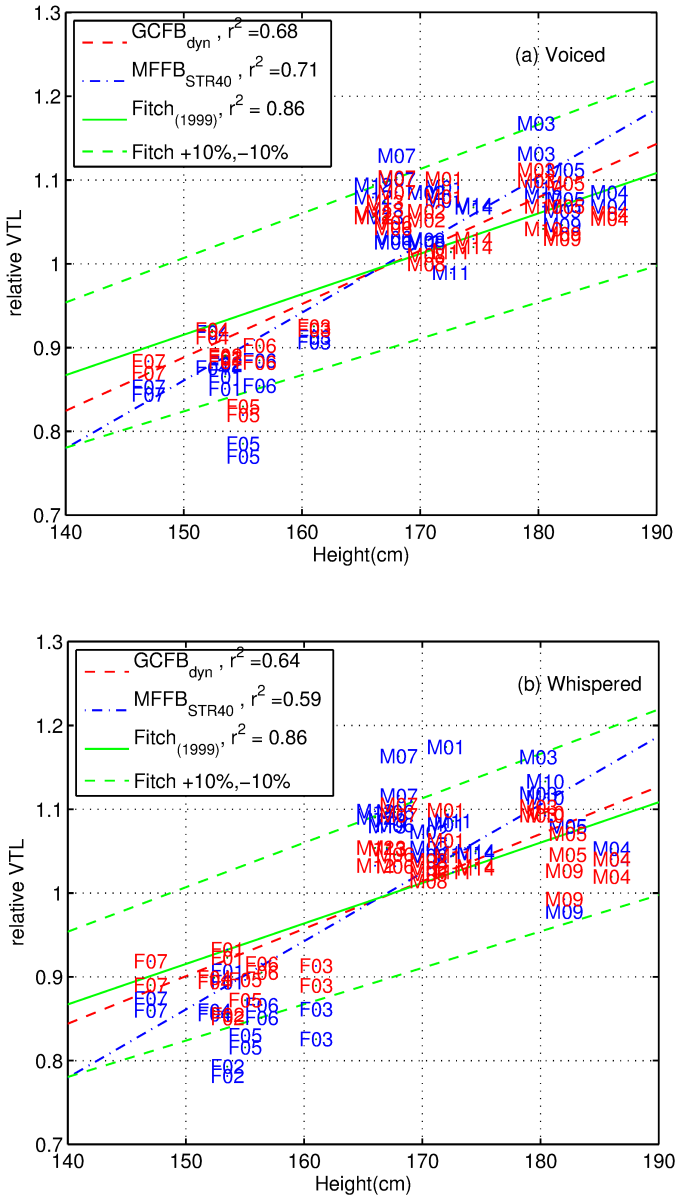

Fig. 11. *Relationship between height and VTL estimated for two sentences from voiced (a) and whispered speech (b). Each label shows speaker ID centered on estimated VTL for one sentence.*

れる.

3.4.6 声道長と身長の関係

前述のとおり使用した音声データベースには話者の身長データ情報もある. そこで, 身長と推定声道長との関係を, Fitch と Giedd [33] による MRI 計測から求めた身長と声道長

の関係と比較することとした. そこでは, 2 歳から 25 歳の間の 121 人の測定から得た, 身長 (Height) から声道長 (VTL) への回帰直線が次式になることが報告されている.

$$VTL = 2.7 + 0.068 \times Height\ (cm), \tag{3.1}$$

ここで, $r = 0.926$(adj.), $r^2 = 0.86$, $p < 0.0001$.

図 11 は, 身長と声道長の関係を, (a) 通常発声の場合, (b) ささやき声の場合についての, 身長に対する推定された声道長 $\hat{l} = [\hat{l_1}, \hat{l_2}, ..., \hat{l_{28}}]$ の関係の散布図である. ラベルは発話者の ID で, M は男性, F は女性で, 赤文字は $GCFB_{dyn}$, 青字は $MFFB_{STR40}$ を用いて推定された結果である. 同じラベルは 2 文章から求めた値をそれぞれ示している. Fitch と Giedd による MRI データ [33] からの回帰直線 (緑実線, $r^2 = 0.86$) とその ±10% 区間 (緑破線) も示してある. $GCFB_{dyn}$ による推定結果の回帰直線の決定係数は $r^2 = 0.64$ で, MRI データよりは低いが $MFFB_{STR40}$ の $r^2 = 0.59$ よりも若干高いという結果になった. ただし, 今回の推定は 21 名という少数サンプルで, しかも同性間では身長差がそれほどない大人だけのデータであるので, ばらつきが相対的に大きい. これは年齢とともに成長曲線に沿って身長が伸びる子供も含むこの MRI データとは異なる点で, 決定係数で単純には比較できない. むしろ, 推定の外れ値や発話文章ごとの違いがどの程度あるかで評価した方が良い可能性がある. $GCFB_{dyn}$ を用いると, 一話者 (F05) の通常発声 (図 11 (a)) 以外はすべて ±10% 区間 (緑破線) 内に推定されていることがわかる. しかし, $MFFB_{STR40}$ の場合, 特にささやき声 (図 11 (b)) で外れ値が多い. また, それに伴い, 文章ごとの違いも大きくなっていることがわかった. したがって, $GCFB_{dyn}$ を用いる方が、$MFFB_{STR40}$ よりも、安定な推定をできることがわかった.

3.5 声道長推定のまとめ

GCFB 聴覚フィルタバンクを用いると, 一般に広く使われて来たメル周波数フィルタバンク MFFB を用いるよりも安定して頑健な推定ができることがわかった. 図 7 から, $MFFB_{STR40}$ は STRAIGHT スペクトルを用いており, フーリエ変換を用いた従来法の $MFFB_{STFT24}$ よりも改良されて良くなっていが, 限界があるようである. $GCFB_{dyn}$ は, 線形の $GCFB_{lin}$ よりも良いこともわかった. $GCFB_{dyn}$ 外界の音圧依存で時間的に変動する非線形フィルタバンクで, 線形のウェーブレット変換よりも複雑である. しかし, 基本構成がウェーブレット変換となっていて, その上に制約のある非線形性が乗っている形になっている. 次節でこのことについて紹介する.

4. 線形のウェーブレット変換を超えて

スケール変形に対して理論的に最適なはずの線形のウェーブレット変換よりも, 非線形の聴覚フィルタバンクの方が安定に声道長推定できることがわかった. このことは, この声道長推定の問題が単純なスケール変形だけでは表すことができないということを示して

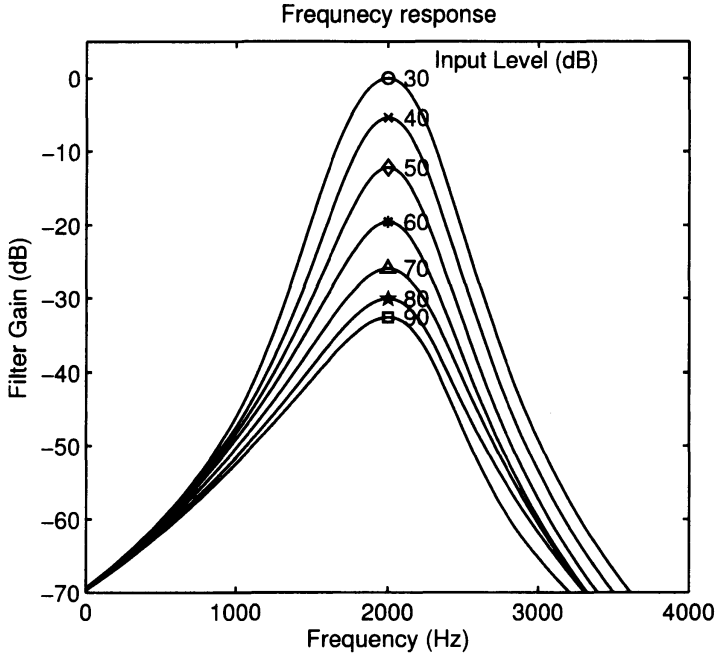

Fig. 12. Level dependent gain and filter shape when the input sound pressure level is varied between 30 and 80 dB.

いる. ここでは, 聴覚末梢系の非線形特性を反映させているガンマチャープ聴覚フィルタバンク（$\mathrm{GCFB_{dyn}}$）の非線形性について紹介し, 理論構築の議論の導入としたい.

4.1 周波数範囲

前節の結果は, 安定な声道長推定には 500Hz 以上の周波数領域を用いることが重要であるということを示している. 音声において声道長のスケール性だけが表出するのであれば, スケール変形に対し「透明」なはずのウェーブレット変換を用いれば十分で, 周波数を制約する条件は出ないはずである. しかしながら, 音声を駆動するための声帯振動があるため, その基本周波数 F_0 と高調波の影響がどうしても出てくる. また, 聴覚末梢系の特性に関しても, 図 4 に示したように, 定 Q 型フィルタとなるのは, 500Hz 以上である. この下限周波数が一致するのが偶然なのか必然性があるのかは, 今後の検討を待つ必要がある.

4.2 非線形性の効用

図 7 の結果から, 線形フィルタバンク ($\mathrm{GCFB_{lin}}$ ～$\mathrm{MFFB_{STFT120}}$) に対して, 聴覚末梢系の非線形特性を反映させた $\mathrm{GCFB_{dyn}}$ を用いた方が推定精度が良いことがわかる.

図 12 に, 心理物理実験によって求められたガンマチャープ聴覚フィルタの入力音圧に対するフィルタの振幅周波数特性の変化を示す [20]. まず, 音圧が高くなるにつれて, 中心

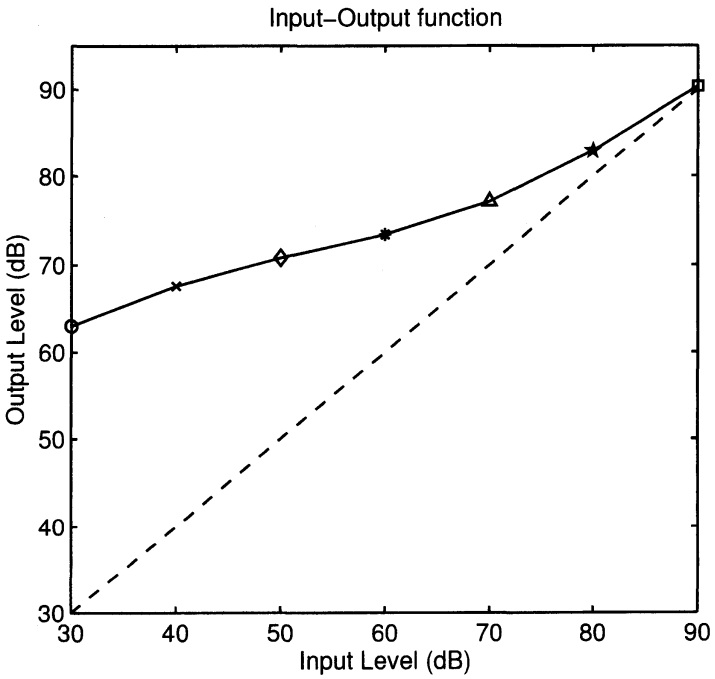

Fig. 13. Input-output function of auditory filter. The solid line shows compressive characteristics with growth rate of 0.2 ～ 0.3 dB/dB.

周波数 (2000 Hz) におけるフィルタの利得が減少することがわかる. また, 中心周波数よりも離れた周波数 (例えば 1000 Hz 以下や 3000Hz 以上) では, レベル依存性がほとんど無いこともわかる. また, この特性を実現するフィルタにおいて, インパルス応答における瞬時周波数変化がほとんど無いことも生理学的にも知られており, ガンマチャープ聴覚フィルタにもその特性を反映させている [22].

この聴覚フィルタに入力した音の音圧レベルと, 基底膜振動の振動のレベルの関係を取ると図 13 に示すような入出力関数になる. 縦軸, 横軸とも dB 値で, 破線の対角線が入出力が 1:1 の線形の場合である. 健聴者の聴覚フィルタにおいては, 入力音圧の増加に対して出力レベルの増加の割合が少なく, おおよそ 0.2 ～0.3 dB/dB の増加率と考えられている. この増加率が 1 よりも小さいため, 圧縮特性と呼ならわされている. これは, 音圧が低い音を聞こえるだけの振動レベル範囲に増幅する作用を入出力関数から表現していることになる. これが聴覚末梢系の最も大きな非線形性で, 難聴者では, 増幅特性が劣化し, 小さい音が聞き取りにくくなる場合もある. この場合, 圧縮特性で見ると傾きが大きくなっていることになる.

この他にも聴覚末梢系の主な非線形性として, 2 音抑圧 [*6] が知られている. フィルタバ

[*6] 中心周波数に正弦波を入れて観測した場合よりも, さらにその周辺の周波数に 2 つ目の正弦波を加えた場合の方が出力が減少する現象. 入力を増やしたにも関わらず出力が減少する.

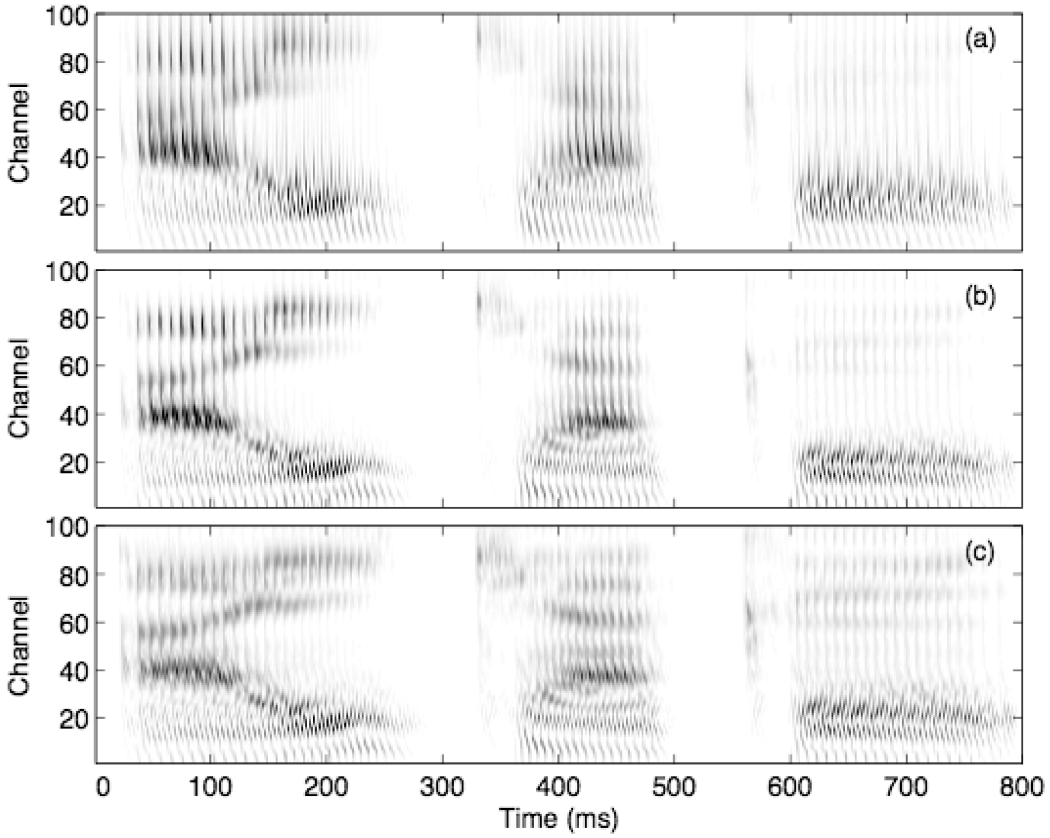

Fig. 14. Cochlear spectrograms, or cochleograms, for the Japanese word 'aikyaku,' plotted on a linear scale to reveal level differences: (a) $GCFB_{partial}$, (b)$GCFB_{lin}$, and (c) $GCFB_{dyn}$.

ンクで末梢系をモデル化する場合に考慮すべき特性である.

これらの非線形性ど導入したガンマチャープ聴覚フィルタバンク $GCFB_{dyn}$ で音声を分析すると,「聴覚的スペクトログラム」を得ることができる. 通常発話の音声「あいきゃく」を分析した例を, 図 14(c) に示す. 同図 (a),(b) は, 線形フィルタバンクの例である. 特に 600 ～ 800 (ms) の所の 80ch 周辺におけるホルマント（声道音響管の共振特性）が強調されて表現されていることがわかる. また、40 ～ 120 (ms) の 40ch 付近のホルマントは線形の場合に比べてむしろコントラストが小さくなっている. このことから, 聴覚フィルタにおける非線形性は, 音声の特徴を最も表す部分を平均的に強調するように働いていることがわかる. これが, 今回の声道長推定においても有効に働いたものと考えられる.

5. おわりに

本論文では，まず寸法知覚（スケール変形の知覚）に関連する初期聴覚系の計算理論の知見を紹介した．聴覚末梢系のモデルであるガンマチャープ聴覚フィルタバンクに関して，スケールを正規化する Mellin 変換の張る空間における最小不確定性から解析的に求められたことや，ウェーブレット性と非線形性があることを述べた，通常発話とささやき声からの声道長推定の問題に関して，他の線形フィルタバンクと対比させ，有利であることを示した．線形の音響管のスケーリング処理に対しては、線形のウェーブレット変換が理論的に最適なはずである。しかし、音声からの声道長推定の場合には，この制約付きの非線形性が有利に働くことを示した．しかしまだ，線形のウェーブレット理論を拡張して，最適性を示すには至っていない．今後の展開を期待したい．

謝辞 本研究の一部は，科学研究費補助金課題番号 19200017, 21300069, 25280063 による支援を受けた．声道長推定に関しては岡本恵里香氏の研究によって進展した．ここに感謝する．

Appendix A. ガンマチャープ関数の導出

ガンマチャープ関数は Mellin 変換が張る空間の最小不確定性を持つ関数として求めることができる [21].

A.1 Mellin 変換

信号 $s(t), (t > 0)$ のメリン変換 [35] は

$$S(p) = \int_0^\infty s(t)t^{p-1}dt,$$

ここで p は複素変数である. 重要な特徴として,

$$\text{if } s(t) \Rightarrow S(p), \quad \text{then } s(at) \Rightarrow a^{-p}S(p),$$

が成立する. ここで矢印は変換を示し, a は実数の伸縮 (スケール) 係数である. すなわち, スケール変形に対し正規化した $S(p)$ の絶対値分布は変化せず, スケール不変表現となる.

A.2 演算子法と Mellin 変換

量子力学において, アフィン変数を使ってスケール性を議論することは既におこなわれている [36]. また, 信号処理の時間周波数表現において, 量子力学で用いられてきた演算子法が表現の類似性から導入されている [37]. 時間演算子 $\mathcal{T} = t$, 時間領域における周波数演算子 $\mathcal{W} = -j(d/dt)$ を導入する. すると Cohen による「スケール演算子」は,

$$C = \frac{1}{2}(\mathcal{TW} + \mathcal{WT}) = \mathcal{TW} - \frac{1}{2}j,$$

と表される. これは, 量子力学におけるアフィン変数を表現する演算子として既に知られている [36]. この演算子に対応する「スケール変換」[37] は

$$D(c) = \frac{1}{\sqrt{2\pi}} \int_0^\infty s(t)t^{-jc-1/2}dt,$$

で, メリン変換において $p = -jc + 1/2$ と設定したものである. この式の適用範囲を広げるためバイアス項の実数 c_0 と μ を導入して,

$$p = -j(c - c_0) + (\mu + 1/2).$$

と拡張する. 対応するメリン演算子は次式となる.

$$C_m = \mathcal{TW} + \{c_0 + j(\mu - 1/2)\}.$$

我々の関心事は, 聴覚末梢系のフィルタバンク表現である. そこで, 「周波数シフト項」ω_0 を各々のフィルタを特定するために導入する. すると演算子は以下のように変形できる.

$$C_a = \mathcal{T}(\mathcal{W} - \omega_0) + \{c_0 + j(\mu - 1/2)\}.$$

この周波数シフト項 ω_0 は, 本文の 2.2.3 節で述べた寸法形状イメージ (SSI) における周波数正規化の処理により, 完全に取り除くことができる. このため, メリン変換の枠組みからは外れない. 時間とこの演算子の交換子は以下となる.

$$[\mathcal{T}, C_a] = [\mathcal{T}, C_m] = [\mathcal{T}, C] = j\mathcal{T}.$$

交換子が 0 とならないので, 時間とこの演算子の表すメリン空間の値は独立に計測できない. この時の不確定性の関係は, 以下で表される.

$$\Delta t \cdot \Delta c_a \geq \frac{1}{2}| < [\mathcal{T}, C_a] > | = \frac{1}{2}| < j\mathcal{T} > | = \frac{< t >}{2}.$$

ここで, Δ. は 標準偏差, $< . >$ は平均を表す. 関数の時間平均値の 1/2 以上という条件となる. 次節で, この最小不確定性を満たす関数を導出する.

なお, 良く知られている時間-周波数空間における不確定性の関係は次式である [37].

$$\Delta t \cdot \Delta \omega \geq \frac{1}{2}| < [\mathcal{T}, \mathcal{W}] > | = \frac{1}{2}| < j > | = \frac{1}{2}.$$

この最小値を取るのは, もちろん Gabor 関数 [38] である.

A.3 最小不確定性を満たす関数

演算子が定義できると, 最小不確定性を持つ関数は固有値問題を解くことによって求められる. C_a や C_m は, $\mu = 0$ の場合以外 Hermitian ではない. しかし, 平均値を引いた $(C_a - < C_a >)$ は Hermitian となるため実固有値が求まる. この演算子と時間とで張る空間における最小不確定性を持つ関数は, 以下の固有値問題の解として得られる.

$$(C_a - < C_a >)\, s(t) = \lambda(\mathcal{T} - < t >)\, s(t).$$

ここで

$$\lambda = \frac{< [\mathcal{T}, C_a] >}{2(\Delta\mathcal{T})^2} = \frac{j < t >}{2(\Delta t)^2},$$

である. 固有値問題の式を展開すると以下のようになる.

$$t(-j\frac{d}{dt})\, s(t) - (\omega_0 + j\alpha_1)t\, s(t) + (-c_1 + j\alpha_2)\, s(t) = 0.$$

ここで, $\alpha_1 = < t > /2(\Delta t)^2$, $\alpha_2 = \mu - 1/2 - \mathrm{Im} < c_a > + < t >^2 /2(\Delta t)^2$, $c_1 = \mathrm{Re} < c_a > - c_0$ で, Re., Im. はそれぞれ実部, 虚部を示す. この解は, 以下のように求まる.

$$\begin{aligned} s(t) &= a\, t^{\alpha_2 + jc_1} \exp(-\alpha_1 t + j\omega_0 t), \\ &= a\, t^{\alpha_2} \exp(-\alpha_1 t) \exp(j\omega_0 t + jc_1 \ln t). \end{aligned}$$

ここで a は定数で, ln は自然対数である.

この包絡線 $t^{\alpha_2} \exp(-\alpha_1 t)$ はガンマ分布関数 $\gamma(t)$ である. 搬送波は $\exp(j\omega_0 t + jc_1 \ln t)$ で表される. 搬送波の偏角を時間微分すると, 瞬時周波数 f_i が得られる.

$$f_i = \frac{1}{2\pi}(\omega_0 + \frac{c_1}{t}).$$

これは, 時間的に瞬時周波数が変化することを示しており, 音として再生するとチャープ音である. そこで, この関数を「ガンマチャープ (gammachirp)」と命名した. ここで, $c_1 = 0$ とすると, 搬送波は一定周波数の正弦波となり,「ガンマトーン (gammatone)」関数となる. すなわち, ガンマチャープは, 元々実験式として与えられたガンマトーンを特殊解として持つ, 自然な形の拡張となっていることがわかる.

Appendix B. 声道長推定手法

ここでは, 文献 [3, 12–16] における声道長推定法について簡単に紹介する.

B.1 スペクトル距離に基づく声道長比の推定

同じ文章を発話した話者 A, 話者 B の音声はフィルタバンクによって分析され, 平滑化されたスペクトログラム $P_A(\tilde{f}, t)$ と $P_B(\tilde{f}, t)$ が求められる. ここで $\tilde{f}$ はワープ周波数で, フィルタバンクにより ERB 周波数 f_{ERB} あるいは mel 周波数 f_{mel} のいずれかを表す. また,t は分析時刻 (分析窓の中心時刻) を表す. 二つの音声の音素の出現位置は異なっているため, まず,B のスペクトログラムの時間軸を A と合うように変形する. 変形したスペクトログラムは $P_{Bn}(\tilde{f}, t)$ で表される. A と B の声道長を一致させるために, $P_{Bn}(\tilde{f}, t)$ を, 元の周波数軸上で線形に r 倍伸縮させる [*7]. 周波数は r 倍されて rf となり, これをワープ周波数に変換すると $\tilde{rf}$ となる. したがって, 変形されたスペクトルは $P_{Bn}(\tilde{rf}, t)$ で表される. 分析時刻 t のとき, dB 上でのスペクトル距離は, 実効値 (rms) として以下の式で表される.

[*7] 単純な音響管近似で伸縮が行われていると仮定している. 実際の関係はもう少し複雑で 1 次関数以上が必要な可能性もある. しかし, この分野で十分な検討はまだ行われていない. そこでは, データのばらつきとフィッティングの良さのトレードオフや, 説明変数の少なさ (オッカムの剃刀) も含む AIC 規準等を適用して考えるべきであろう.

(Appendix B.1)
$$D_{dB}(t,r) = \sqrt{\frac{D_P}{\tilde{f}_H - \tilde{f}_L}} \quad ,$$

where

$$D_P = \int_{\tilde{f}_L}^{\tilde{f}_H} \left(10\log_{10} \frac{P_A(\tilde{f},t)}{\bar{P}_A(t)} - 10\log_{10} \frac{P_{Bn}((r\tilde{f}),t)}{\bar{P}_{Bn}(t)} \right)^2 d\tilde{f},$$

$\tilde{f}_L$ と $\tilde{f}_H$ は周波数帯域の下限周波数と上限周波数, $\bar{P}_A(t)$ と $\bar{P}_{Bn}(t)$ は周波数の平均値である.

最適な声道長比の推定値 r は, 文章全体の距離 D_{dB}^{total} を最小化する値である.

(Appendix B.2)
$$r = \mathrm{argmin}(D_{dB}^{total}(r)),$$

where total distance $D_{dB}^{total}(r)$ is defined by using frame-wise spectral distance $D_{dB}(t,r)$ in Eq. Appendix B.1:

(Appendix B.3)
$$D_{dB}^{total}(r) = \sqrt{\frac{1}{T}\int_0^T D_{dB}^2(t,r)dt},$$

ここで、T は、フレーム処理の最終フレームに相当する. これにより, 有声・無声にかかわらず, 文全体を処理することができる.

B.2 声道長比の推定手法

図 5 に, その計算方法の概略を示す. i 番目と j 番目の話者の VTL を l_i と l_j, とすると, 声道長比は $r_{i,j} = l_i/l_j$ で表される. 対数をとることで, 差分で表すことができる.

$$\log(r_{i,j}) = \log(l_i) - \log(l_j).$$

声道長比を求める際, フィルタバンク上で伸縮処理が片一方だけに適用されて, 式の上でバランスが取れない. そこで, 単純に 2 つの組合せではなく逆順も考慮した順列とした (cf. [3]). たとえば, 28 名から 2 名ずつ並べる順列は 756($=_{28}P_{27}$) 通りである.

$$\begin{bmatrix} \log(r_{1,2}) \\ \log(r_{1,3}) \\ \vdots \\ \log(r_{27,28}) \\ \log(r_{2,1}) \\ \log(r_{3,1}) \\ \vdots \\ \log(r_{28,27}) \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ -1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix} \begin{bmatrix} \log(l_1) \\ \log(l_2) \\ \vdots \\ \log(l_{28}) \end{bmatrix}$$

ここで, 最後の行は, 声道長の幾何学的平均値を正規化する制約で, 行列を正則化するために導入している. 実際の声道長の情報が無いため, 声道長 l_i はお互いの相対値として求まる.

上式の左辺を $\boldsymbol{r}_{log}(=\log(\boldsymbol{r}))$, 右辺の係数行列を H, 声道長の対数のベクトルを $\boldsymbol{l}_{log}(=\log(\boldsymbol{l}))$ と書き直す.

$$\boldsymbol{r}_{log} = H\boldsymbol{l}_{log}.$$

ここで最小 2 乗近似を行い, 声道長 $\hat{\boldsymbol{l}} = [\hat{l_1}, \hat{l_2}, ..., \hat{l_{28}}]$ の推定値を計算する.

$$\hat{\boldsymbol{l}}_{log} = (H^T H)^{-1} H^T \boldsymbol{r}_{log},$$
$$\hat{\boldsymbol{l}} = [\hat{l}_1, \hat{l}_2, ..., \hat{l}_{28}]^T = \exp(\hat{\boldsymbol{l}}_{log}).$$

また, $\hat{\boldsymbol{l}}_{log}$ から声道長比の推定値 $\hat{\boldsymbol{r}}$ も計算できる.

$$\hat{\boldsymbol{r}} = \exp(H\hat{\boldsymbol{l}}_{log}).$$

この値と, スペクトル距離から計算した声道長比 $\boldsymbol{r}$ との間のユークリッド距離 d_{est} (rms 値) で, 推定誤差を評価できる.

(Appendix B.4) $$d_{est} = \|\boldsymbol{r} - \hat{\boldsymbol{r}}\| \simeq \sigma.$$

d_{est} は, 図 6 の恒等写像線 ($\hat{\boldsymbol{r}} = \boldsymbol{r}$) を中心とした標準偏差 σ とほぼ同じである. したがって σ が小さいほど, 組み合わせの条件の相違による変動が小さく, 安定に推定できていると考えることができる.

B.2.1 周波数領域の選択

安定な声道長推定のためには, 式 Appendix B.1 中の周波数領域 $[f_L, f_H]$ を適切に選ぶ必要がある. まず, スペクトルの低い周波数範囲には, 声帯の振動の速度や波形による影響が大きく出る. 発話に伴い動的に大きく変化する基本周波数 ($F0$) の成分が主要なスペクトル上のピークを形成し, 声道の共鳴のピークとは異なる. また, 高い周波数領域においては, 個人ごとに異なる梨状窩 [31] の共鳴による影響が出る. 中間の周波数では, これらの影響を受けにくく声道によるスペクトルピーク情報が最も強くなるため, 声道長を効果的に推定することができるものと考えられる. すなわち式 Appendix B.2 の $\boldsymbol{r}$ は, 周波数領域 $[f_L, f_H]$ の関数となる. したがって様々な $[f_L, f_H]$ の組み合わせにおいて $\boldsymbol{r}$ を求め, その中で式 Appendix B.4 の d_{est} を最小にする 最良周波数領域 $[f_L, f_H]$ を選ぶことが最終的な目的となる.

参考文献

[1] Wakita, H., "Normalization of vowels by vocal-tract length and its application to vowel identification, " IEEE Trans. Acoust., Speech, Signal Processing, ASSP-32, pp. 183–192,1977.

[2] 浅香佳希, 西田沙織, 赤桐隼人, 西村竜一, 入野俊夫, 河原英紀, "声道長の正規化に基づく簡易モーフィング音声の品質改良について," 信学会音声研究会, SP2009-34, 109(99), pp.63–68, 2009.

[3] Okamoto, E, Irino, T., Nisimura, R, Kawahara, H., "Evaluation of voice morphing using vocal tract length normalization based on auditory filterbank," Proc. NCSP'11, pp.187–190, 2011.

[4] Irino, T. and Patterson, R.D., "Segregation information about the size and shape of vocal tract using a time-domain auditory model : The stablished wavelet-Mellin transform," Speech Communication, 36(3-4), pp.181–203, 2002.

[5] 入野俊夫,"音源の形状情報と寸法情報を分離する聴覚でのイメージング," 日本音響学会誌, 56 巻 7 号, pp. 505–508, July 2000.

[6] 入野俊夫, "初期聴覚系におけるスケール理論," 音響学会春季研究発表会講演論文集 I, pp.511–514, 2003.

[7] Patterson, R.D., "Auditory images: How complex sounds are represented in the auditory system," J. Acoust. Soc. Japan (E), 21, pp. 183–190, ,2000 (入野抄訳 , "聴覚イメージ : 複雑な音が聴 覚システムでいかに表現されるか," 日本音響学会誌, 56 巻 7 号, pp. 503-504, July 2000.)

[8] Smith, D.R., Patterson, R.D. , Turner, R., Kawahara, H. and Irino, T., "The processing and perception of size information in speech sounds," J. Acoust. Soc. Am, 117(1), pp.305–318, 2005.

[9] 青木良枝, 入野俊夫, パターソン ロイ, 河原英紀, "スケール変形した有声／無声単語の寸法弁別と音韻認識に関する検討," 日本音響学会聴覚研究会資料, H-2008-89, Vol. 38, No. 5, pp.507–512, 2008.

[10] Irino, T., Aoki, Y., Kawahara, H., and Patterson, R.D., "Size Perception for acoustically scaled sounds of naturally pronounced and whispered words," in "Neurophysiological Bases of Auditory Perception," Enrique A. Lopez-Poveda, Alan R. Palmer, and Ray Meddis (Eds.), pp.235–243, Springer, LaVergne, TN USA, 644p., Apr., 2010.

[11] Irino, T., Aoki, Y., Kawahara, H., and Patterson, R.D., "Comparison of performance with voiced and whispered speech in word recognition and mean-formant-frequency

discrimination," Speech Commun., 54 (9), pp.998-1013, 2012.

[12] 岡本恵里香, 西村竜一, 入野俊夫, 河原英紀, "聴覚フィルタバンクを用いた声道長比推定, " 電子情報通信学会 音声研究会, 電子情報通信学会技術研究報告, Vol.111, No.153, SP2011-43, pp.11–16, 2011 年 7 月.

[13] Okamoto, E., Irino, T., Nisimura, R., Kawahara, H., "Auditory filterbank improves voice morphing," in Proc. Interspeech 2011, Tue-Ses2-P1, Florence, Italy, 27-31 Aug., 2011.

[14] 岡本恵里香, 西村竜一, 入野俊夫, 河原英紀, "聴覚フィルタバンクを用いた声道長推定法の比較, " 日本音響学会：春季研究発表会講演論文集, 3-Q-15, 2011 年 9 月.

[15] 岡本恵里香, 北出晴香, 西村竜一, 河原英紀, 入野俊夫, "聴覚フィルタバンクによる声道長推定と身長との相関および発話様式の影響," 日本音響学会聴覚研究会資料, Vol.42, No.1, H-2012-7, pp.35-40, 2012.

[16] Irino, T., Okamoto, E., Nisimura, R., and Kawahara, H., "Vocal tract length estimation for voiced and whispered speech using Gammachirp Filterbank," Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference (APISPA ASC 2013), OS13-SLA.5-5, #138, Kaohsiung, Taiwan, 29 Oct. - 1 Nov. 2013.

[17] Moore, B. C. J., "Psychology of Hearing (5th ed)," Academic Press, London, 2003. (大串訳「聴覚心理学概論（第 3 版)」誠信書房)

[18] 入野俊夫, "はじめての聴覚フィルタ," 音響学会誌, 66（10）, pp. 506-512, 2010.

[19] 入野俊夫, "はじめての聴覚フィルタ - 心理物理実験デモで学ぶ聴覚フィルタ特性 -," 秋季音講論, pp.1347 –1348, 2010.

[20] 入野俊夫, "聴覚フィルタの心理物理実験とモデル（第 4 章)" , in "聴覚モデル" (森, 香田編著), p.233, pp.101–128, コロナ社, 東京 , 2011.

[21] Toshio Irino and Roy D. Patterson "A time-domain, level-dependent auditory filter: the gammachirp," J. Acoust. Soc. Am., 101 (1), pp.412–419, 1997.

[22] Irino, T. and Patterson, R. D., "A compressive gammachirp auditory filter for both physiological and psychophysical data," J. Acoust. Soc. Am., 109 (5), pp.2008-2022, May 2001.

[23] Irino, T. and Patterson, R. D., "A dynamic compressive gammachirp auditory filterbank," IEEE Trans. Audio, Speech, Lang. Process., 14(6), pp. 2222–2232, Nov. 2006.

[24] Daubechies, I., "The wavelet transform, time-frequency localization and signal analysis ," IEEE Trans. Information Theory, Vol. 36 (5), pp. 961–1005, 1990.

[25] Patterson, R. D., Allerhand, M. and Giguére, C., "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform, " J. Acoust.

Soc. Amer., vol. 98, pp. 1890–1894,1995.

[26] de Boer, E. and de Jongh, H.R., "On cochlear encoding: Potentialities and limitations of the reverse-correlation technique," J. Acoust. Soc. Am., 63, pp. 115–135, 1978.

[27] Patterson, R.D., Unoki, M., Irino, T.,"Extending the domain of center frequencies for the compressive gammachirp auditory filter," J. Acoust. Soc. Amer., vol. 114 (3), pp. 1529–1542, 2003.

[28] Lyon, R.F, Ponte, J., and Chechik, G., "Sparse coding of auditory features for machine hearing, " ICASSP2011, 2011.

[29] Versnel, H. and Shamma S.A., "Spectral-ripple representation of steady-state vowels in primary auditory cortex, " J. Acoust. Soc. Am., 103(5), pp. 2502–2514, 1998.

[30] 津崎実, 入野俊夫,"シミュレータによる内部表現と特徴量（第 7 章）," in "聴覚モデル" (森, 香田編著) p.233, pp.195–229, コロナ社, 東京 , 2011.

[31] Dang, J. and Honda, K. ,"Acoustic characteristics of the piriform fossa in models and humans, " *J. Acoust. Soc. Am.*, 101(1), pp. 456–465, 1997.

[32] http://labrosa.ee.columbia.edu/matlab/rastamat/ の HTK のメル尺度を選択 (最終アクセス日：24 Apr 2014)

[33] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *J. Acoust. Soc. Amer.*, 106(3), pp. 1511–1522, 1999.

[34] Atake, Y., Irino, T., Kawahara, H., Lu, J., Nakamura, S. and Shikano, K., "Robust fundamental frequency estimation using instantaneous frequencies of harmonic components," 6th International Conference on Spoken Language Processing, ICSLP2000, No.867 , Beijing, China, 2000.

[35] Titchmarsh, E. C., "Introduction to the Theory of Fourier Integrals," Oxford U.P., London, 2nd ed, 1948.

[36] Klauder, J. R., "Path integrals for affine variables," in Functional Integration: Theory and Applications, edited by Antoine, J. P. and Tirapgui, E. , Plenum, New York, 1980.

[37] Cohen, L. "The scale representation," IEEE Trans. Signal Process. 41, pp. 3275–3292, 1993.

[38] Gabor, D., "Theory of communication," J. IEE (London), 93, pp. 429–457, 1946.