Vol.2016-MUS-111 No.20
2016/5/21

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

山本 克彦[1,a)]　入野 俊夫[1,b)]　松井 淑恵[1,c)]　荒木 章子[2,d)]　木下 慶介[2,e)]　中谷 智広[2,f)]

概要：聴覚モデルベースで音声明瞭度を予測する客観的評価指標は，音声強調処理技術を評価するためにも必要不可欠である．しかし，従来法 (sEPSM) では聴覚末梢系の音圧依存特性（圧縮特性）が反映されていない上に，スペクトル減算法以外の非線形な音声強調処理手法では評価されていなかった．本研究では，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の予測法 (dcGC-sEPSM) を提案した．非線形の音声強調処理手法であるスペクトル減算法とウィナーフィルタ型の雑音抑圧法を用いて聴取実験を行った．客観的評価として，提案法 (dcGC-sEPSM) および既存法 (GT-sEPSM, CSII, STOI) を用いて音声明瞭度の予測を行った．聴取実験から得られた音声明瞭度と比較した結果，提案法は既存法よりも聴取実験の傾向に近い音声明瞭度を予測することがわかった．

Predicting speech intelligibility using the dynamic compressive gammachirp filterbank: comparison with the result for enhanced speech

Yamamoto Katsuhiko[1,a)]　Irino Toshio[1,b)]　Matsui Toshie[1,c)]　Araki Shoko[2,d)]
Kinoshita Keisuke[2,e)]　Nakatani Tomohiro[2,f)]

Abstract: An objective measure index of speech intelligibility based on auditory models is essential to evaluate speech enhancement techniques. The conventional method (sEPSM) was proposed to account for subjective results on a spectral subtraction, but has not been tested by recent state-of-the-art speech enhancement algorithms. We developed a new method using the dynamic compressive gammachirp auditory filterbank (dcGC-sEPSM) for speech intelligibility (SI) prediction of synthetic sounds processed by nonlinear speech enhancement algorithms. Subjective experiments were performed by using the spectral subtraction and a recent Wiener filter algorithm. We compared the subjective SI scores with the objective SI scores predicted by the proposed dcGC-sEPSM, the original sEPSM, the three-level coherence SII (CSII), and the short-time objective intelligibility (STOI). The results show that the dcGC-sEPSM performs better than the conventional models.

1. はじめに

雑音中に埋もれた音声情報を強調する雑音抑圧処理手法が多数提案されている．しかし, 聴取実験による音声明瞭度と相関の高い客観評価指標は依然として確立されていない．現在，Speech Intelligibility Index (SII) [1] や Speech Transmission Index (STI) [2] などが音声明瞭度の客観的評価指標として国際標準規格化されている．SII は，電話の帯域や雑音重畳音声の影響を評価するための音声明瞭度指数として作成され，STI は室内音響の影響を評価するための音声伝達指数として作成された．これらの客観的評価指標は室内音響等の線形歪みに対しては有効ではあるが，非線形の音声強調の処理音声等には適用できない．非線形の音声強調の処理音声等に対応するために，様々な拡張手法が提案されている [3–6]．Kates & Arehart [5]

[1] Wakayama University
Wakayama 640–8510, Japan
[2] NTT Communication Science Laboratories
Kyoto 619–0237, Japan
[a)] s149011@sys.wakayama-u.ac.jp
[b)] irino@sys.wakayama-u.ac.jp
[c)] tmatsui@sys.wakayama-u.ac.jp
[d)] shoko.araki@lab.ntt.co.jp
[e)] kinoshita.k@lab.ntt.co.jp
[f)] nakatani.tomohiro@lab.ntt.co.jp

は，SII をピーククリッピングに対応させるように拡張した three-level coherence SII (CSII) を提案した．Taal *et al.* は，非線形な信号処理の一つである ideal time-frequency segregation (ITFS) に対応した short-time objective intelligibility measure (STOI) を提案した．一方で，Jørgensen & Dau [7] は，人間の聴覚システムの観点から音声明瞭度を予測する speech-based envelope power spectrum model (sEPSM) を提案している．sEPSM は非線形な信号処理手法の一つであるスペクトル減算法に対応することができる．しかし，最先端の音声強調処理では評価されておらず，必ずしも予測値が主観評価値と一致しない場合もあった．そこで，我々は，sEPSM に聴覚末梢系の非線形特性を導入して拡張する方法を提案し，更なる改良を進めている [8]．本稿では，聴取実験結果と提案法 (dcGC-sEPSM) および既存手法 (sEPSM, CSII, STOI) による予測値を，最先端の音声強調処理で評価した結果について報告する．

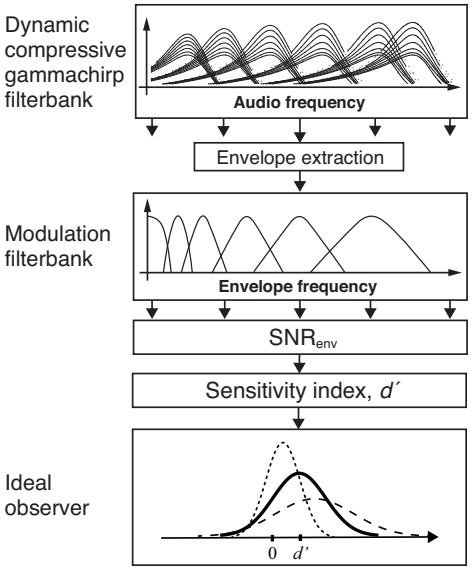


図 1: 提案法 (dcGC-sEPSM) の構成.

2. 提案法

従来法 [7] では，聴覚フィルタバンクに線形のガンマトーン (gammatone; GT) フィルタバンク [9] を用いている．ガンマトーンフィルタは，健聴者の平均的な聴覚特性の第一次近似として提案されたもので，信号レベル依存の周波数選択性や圧縮特性といった聴覚末梢系の非線形性を反映することはできない．そこで，提案法では聴覚フィルタの非線形特性を時々刻々と反映できる動的圧縮型ガンマチャープ (dynamic compressive gammachirp; dcGC) フィルタバンク [10] を用いる．

2.1 動的圧縮型ガンマチャープフィルタバンクによる sEPSM の拡張

図 1 に，従来法 (sEPSM) を動的圧縮型ガンマチャープフィルタバンクで拡張した提案法の構成を示す．以降では，聴覚フィルタバンクにガンマトーンフィルタバンクを用いている従来法を “GT-sEPSM,” 動的圧縮型ガンマチャープフィルタバンクを用いた提案法を “dcGC-sEPSM” と呼ぶ．

GT-sEPSM では，1/3 オクターブ間隔の中心周波数（63 Hz から 8000 Hz）を持つ 22 チャンネルのガンマトーンフィルタバンクを用いて入力信号の周波数分析を行っている．しかし，これでは人間の聴覚系の特性を正確に反映することができないため，本研究では動的圧縮型ガンマチャープフィルタバンク (dcGC-FB) を用いることを提案する．ここで用いた dcGC-FB は，$\mathrm{ERB_N}$ 軸上で等間隔の中心周波数（100 Hz から 6000 Hz）をもつ 100 チャンネルのフィルタから構成される．それ以降の変調周波数領域における信号処理も 100 チャンネルの帯域信号を用いて行う．

2.2 振幅包絡の分析

入力信号は聴覚フィルタバンクで時間応答として出力される．出力された各波形をヒルベルト変換した後，一次の低域通過フィルタ (カットオフ周波数：150 Hz) の出力から振幅包絡が抽出される．これらの振幅包絡の波形全体をフーリエ変換してその周波数特性を分析する．この上で，周波通領域で定義される変調フィルタを重み関数として掛け，変調帯域ごとのパワーを計算する。この変調フィルタバンクは、三次の低域通過フィルタ (カットオフ周波数：1 Hz) と二次の帯域通過フィルタ (チャンネル数 J : 6, 中心周波数：2〜64 Hz, Q 値：1) で構成される．この分析を，音声+雑音 (S+N) と雑音のみ (N) のそれぞれの振幅包絡について行い，各出力のパワースペクトル $P_{env,S+N}$ と $P_{env,N}$ を算出する．なお，強調音声処理後の信号でも同様の処理を行う．

2.3 内部指標 ($\mathbf{SNR_{env}}$) の計算方法

$P_{env,S+N}$ と $P_{env,N}$ を用いて，変調フィルタバンク出力における信号対雑音比を $\mathrm{SNR_{env}}$ を求める．ただし，GT-sEPSM と dcGC-sEPSM では $\mathrm{SNR_{env}}$ の計算方法が異なるため，以下の節において説明する．

2.3.1 GT-sEPSM における計算

GT-sEPSM では，聴覚フィルタバンクおよび変調フィルタバンクからの出力の独立性を仮定しており，以下の式のように聴覚フィルタバンクと変調フィルタバンクのチャンネル毎に $\mathrm{SNR}^{\mathrm{GT}}_{\mathrm{env},i,j}$ を算出する，

$$\mathrm{SNR}^{\mathrm{GT}}_{\mathrm{env},i,j} = \frac{P_{env,S+N,i,j} - P_{env,N,i,j}}{P_{env,N,i,j}}, \tag{1}$$

ここで，$i\{i|1 \le i \le I\}$ はガンマトーンフィルタのチャンネル，$j\{j|1 \le j \le J\}$ は変調フィルタのチャンネルを表す．最終的に，$\mathrm{SNR}^{\mathrm{GT}}_{\mathrm{env}}$ は以下の総和で計算される，

$$\mathrm{SNR}_{\mathrm{env}}^{\mathrm{GT}} = \sqrt{\sum_{j=1}^{J}\sum_{i=1}^{I}\left(\mathrm{SNR}_{\mathrm{env},i,j}^{\mathrm{GT}}\right)^2}. \quad (2)$$

2.3.2 dcGC-sEPSM における計算

動的圧縮型ガンマチャープフィルタバンクを適用した dcGC-sEPSM では聴覚フィルタ間の冗長性を仮定しているため，$\mathrm{SNR}_{\mathrm{env}}$ の計算方法にも変更を加えた．まず，変調フィルタバンク出力のパワースペクトル $P_{env,S+N}$ と $P_{env,N}$ を動的圧縮型ガンマチャープフィルタバンクの全チャンネルで足し合わせる．その後，以下の式のように，変調フィルタバンクのチャンネル毎に $\mathrm{SNR}_{\mathrm{env},j}^{\mathrm{dcGC}}$ を算出する，

$$\mathrm{SNR}_{\mathrm{env},j}^{\mathrm{dcGC}} = \frac{\sum_{i=1}^{I}(P_{env,S+N,i,j} - P_{env,N,i,j})}{\sum_{i=1}^{I} P_{env,N,i,j}}. \quad (3)$$

最後に，$\mathrm{SNR}_{\mathrm{env},j}$ を総和することにより全体の $\mathrm{SNR}_{\mathrm{env}}$ が計算される，

$$\mathrm{SNR}_{\mathrm{env}}^{\mathrm{dcGC}} = \sqrt{\sum_{j=1}^{J}\left(\mathrm{SNR}_{\mathrm{env},j}^{\mathrm{dcGC}}\right)^2}. \quad (4)$$

2.4 音声明瞭度の計算方法

図 1 中の $\mathrm{SNR}_{\mathrm{env}}$ の値は，以下の式で理想観測者 (ideal observer) の感度指標 d' に変換される [7].

$$d' = k \cdot (\mathrm{SNR}_{\mathrm{env}})^q \quad (5)$$

ここで，k と q は定数であり，音声試料や実験条件に左右されないと仮定されている．しかしながら，これらの値は第 2.3 節における $\mathrm{SNR}_{\mathrm{env}}$ の計算方法に大きく依存するため，聴取実験の結果をもとに定数の調整を行った．最終的に，GT-sEPSM ではそれぞれの値を $k = 0.40$ と $q = 0.5$，dcGC-sEPSM では $k = 0.83$ と $q = 0.5$ とした．その後，感度指標 d' を入力として，不等分散ガウスモデル [11] と m 肢強制選択 (mAFC) モデル [12] を反映させた以下の式から音声明瞭度が算出される，

$$P_{correct}(d') = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}}\right). \quad (6)$$

ここで，Φ は累積ガウス分布である．μ_N と σ_N は，音声試料から推測される応答の選択肢の数 m によって決まる．σ_S は，音声試料の冗長性に関連すると仮定したパラメータである．意味のある簡単な文であると σ_S は小さく，冗長性の無い単音節音だと σ_S は大きい．

3. 強調音声の評価

3.1 雑音抑圧処理手法

文献 [7] の Fig. 6 では，単純なスペクトル減算法 [13] に対する評価が行われていた．そこで本研究では，上述のスペクトル減算法に加えて，ウィナーフィルタ型の雑音抑圧処理手法 [14] を用いて評価を行った．以降では，それぞれの処理手法の概要について説明する．

3.1.1 スペクトル減算法 (SS) [13]

雑音が重畳された音声のパワースペクトルから，雑音のみの区間から算出された雑音のパワースペクトルの推定値を減算し，音声だけの振幅スペクトルを推定する．その後，元の位相スペクトルを用いて音を再合成することにより強調音声を作成した．Berouti らの手法 [13] では，減算する度合い (over-subtraction factor, α) を調整することができる．Jørgensen & Dau [7] の報告では α が 1 より増加した際には結果が大きく変化しなかったため，本研究では $\alpha = 1.0$ とした場合について検討した．以降では，この手法を “$\mathrm{SS}^{(1.0)}$” と呼ぶ．

3.1.2 ウィナーフィルタ型雑音抑圧法 ($\mathbf{WF_{PSM}}$) [14]

雑音抑圧法において，音声の確率モデルを用いた手法が近年注目されている．その中でも，音声認識分野において有効性が知られている手法 [14] を対象とした．この手法は，対数メルスペクトル領域においてクリーン音声および雑音のモデルを用いて観測信号の状態を推定する．その後，推定した音声と雑音で設計したウィナーフィルタにて雑音抑圧を行う．対数メルスペクトル領域では，観測信号と推定信号に生じるミスマッチが非線形関数で表されることから，ベクトル Taylor 級数 [15] で線形近似することでモデルパラメータのミスマッチを推定する．Fujimoto らの手法では，雑音の残留度をパラメータ $\{\varepsilon | 0.0 \le \varepsilon \le 1.0\}$ で調整することができる．本研究では，雑音の残留度を $\varepsilon = 0.0$ (雑音を完全に抑圧した状態と仮定), 0.1, 0.2（雑音をある程度残した状態）とした場合について検討した．以降では，この手法をパラメータ値ごとに “$\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$,” “$\mathrm{WF}_{\mathrm{PSM}}^{(0.1)}$,” “$\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$” と呼ぶ．

3.2 聴取実験

音声試料として，親密度別単語了解度試験用音声データセット (FW07) [16] に収録されている男性話者 (mis) の 4 モーラ単語音声を使用した．心的辞書の影響をなるべく排除するために親密度が最も低い音声セット（親密度 1, 20 単語 × 20 グループ）を採用した．音声データは，量子化ビット数を 16 bits，サンプリング周波数を 16 kHz にダウンサンプリングしたものを使用した．音声試料に重畳する雑音としてピンク雑音を使用し，信号対雑音比 (signal-to-noise retio; SNR) を −6 dB から 3 dB の間で 3 dB 毎に変化させた．この雑音重畳音声を原音声として（以降では “Unprocessed” と呼ぶ），3.1.1 節の SS と 3.1.2 節の $\mathrm{WF}_{\mathrm{PSM}}$ の音声強調処理を行った．呈示される音声刺激の総数は，5 種類の条件 (Unprocessed, $\mathrm{SS}^{(1.0)}$, $\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$, $\mathrm{WF}_{\mathrm{PSM}}^{(0.1)}$, $\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$) および 4 種類の SNR ($-6, -3, 0, 3$ dB) から構成される計 400 個とした．なお，単語自体の難易度の違いの効果を打ち消すために，被験者ごとに別々の聴取音声セットを用意した

聴取実験は A 特性等価騒音レベル約 26 dB の聴力実験

用設備（リオン, AT-62W）内で実施した．呈示音は，計算機端末（Apple, Mac mini Late 2012, OS X 10.8.5）に接続したオーディオインターフェース（Fostex, HP-A8）経由でヘッドホン（Sennheiser, HD-580）から出力した．出力の前に，量子化ビット数を 24 bits，サンプリング周波数を 48 kHz にアップサンプリングした．呈示音圧レベルは，人工耳（Brüel & Kjær, Type 4153）と騒音計（Brüel & Kjær, Type 2250-L）を用いて，A 特性平均 65 dB SPL に校正した．

20 歳から 23 歳の男性 4 名と女性 5 名の健聴者が聴取実験に参加した．実験参加者はランダム順に呈示される音声刺激を聴きとり，聴きとった 4 モーラ音声を解答用紙にひらがなで記入した．音声信号の呈示間隔（オンセット）は 4 秒とした．本研究では完全回答のみを正解として，最終的に音声明瞭度を百分率で算出した．全ての実験参加者が，125 Hz から 8000 Hz の範囲のオージオグラムで健聴な聴力なレベルであることを確認した．また，実験に先立ちインフォームドコンセントを実施し，聴取実験の実施に関する同意を得た．

3.3 客観的評価指標による音声明瞭度予測

提案法 (dcGC-sEPSM) と従来法 (GT-sEPSM) を用いた客観的評価指標が，聴取実験の結果を正しく予測できるかを調べるために，被験者ごとに異なる音声セットに対して音声明瞭度を計算した．さらなる比較のために, 既存の客観的評価指標として用いられている three-level coherence SII (CSII) [5] と short-time objective intelligibility measure (STOI) [6] についても検討した．各モデルにおけるパラメータは，予測された音声明瞭度 (Unprocessed) と聴取実験の結果との平均二乗誤差 (mean-squared error; MSE) が最小になるようにフィッティング [17] を行った．

3.3.1 GT-sEPSM と dcGC-sEPSM

従来法と提案法では，式 6 の m と σ_S を音声試料ごとに調整することによって，内部指標（$\mathrm{SNR_{env}}$）が音声明瞭度に変換される．ここで，Amano & Kondo [18] の報告による 4 モーラの心的辞書の大きさの推定値と，今回用いた音声試料の親密度の低さを勘案して，応答の選択肢の数を $m = 20000$ と置いた．フィッティングによって，σ_S の値は，GT-sEPSM では $\sigma_S = 2.85$, dcGC-sEPSM では $\sigma_S = 2.74$ に決定した．

3.3.2 Three-level CSII

Kates & Arehart [5] が提案した three-level coherence SII (CSII) は，強調音声とクリーン音声の相互スペクトル密度から算出される “magnitude-squared coherence (MSC)” と呼ばれる関数を使用する．CSII では，SII で行われる SNR の計算を signal-to-distortion ratio (SDR) に置き換えている．この SDR の値は，16 チャンネルの ro-ex フィルタ [19] を通過した狭帯域音声の FFT スペクトルの MSC から導出される．CSII の値は，窓長 30 ms のハニング窓 (75%オーバーラップ) の短時間フレームごとに計算され，フレーム間の狭帯域音声の音圧レベルから 3 つのレベルに分類される．その後，それぞれの 3 種類の CSII の値 ($\mathrm{CSII_{low}}$, $\mathrm{CSII_{mid}}$, $\mathrm{CSII_{high}}$) に重み付けを行い，以下のロジスティック関数を用いて強調音声の音声明瞭度を予測する，

$$c = -2.63 - 9.40\mathrm{CSII_{low}} + 11.33\mathrm{CSII_{mid}} - 0.01\mathrm{CSII_{high}}, \quad (7)$$

$$P_{correct}^{(CSII)} = \frac{100}{1 + e^{-c}}. \quad (8)$$

3.3.3 STOI

Taal *et al.* [6] によって提案された short-time objective intelligibility measure (STOI) では，1/3 オクターブバンドを通過した狭帯域音声 (クリーン音声と強調音声) の振幅包絡間の相関係数を用いる．相関係数の値は 384 ms の短時間フレームごとに全チャンネルで計算され，全ての値を平均化する．最終的に，この値は以下のロジスティック関数を用いて音声明瞭度に変換される，

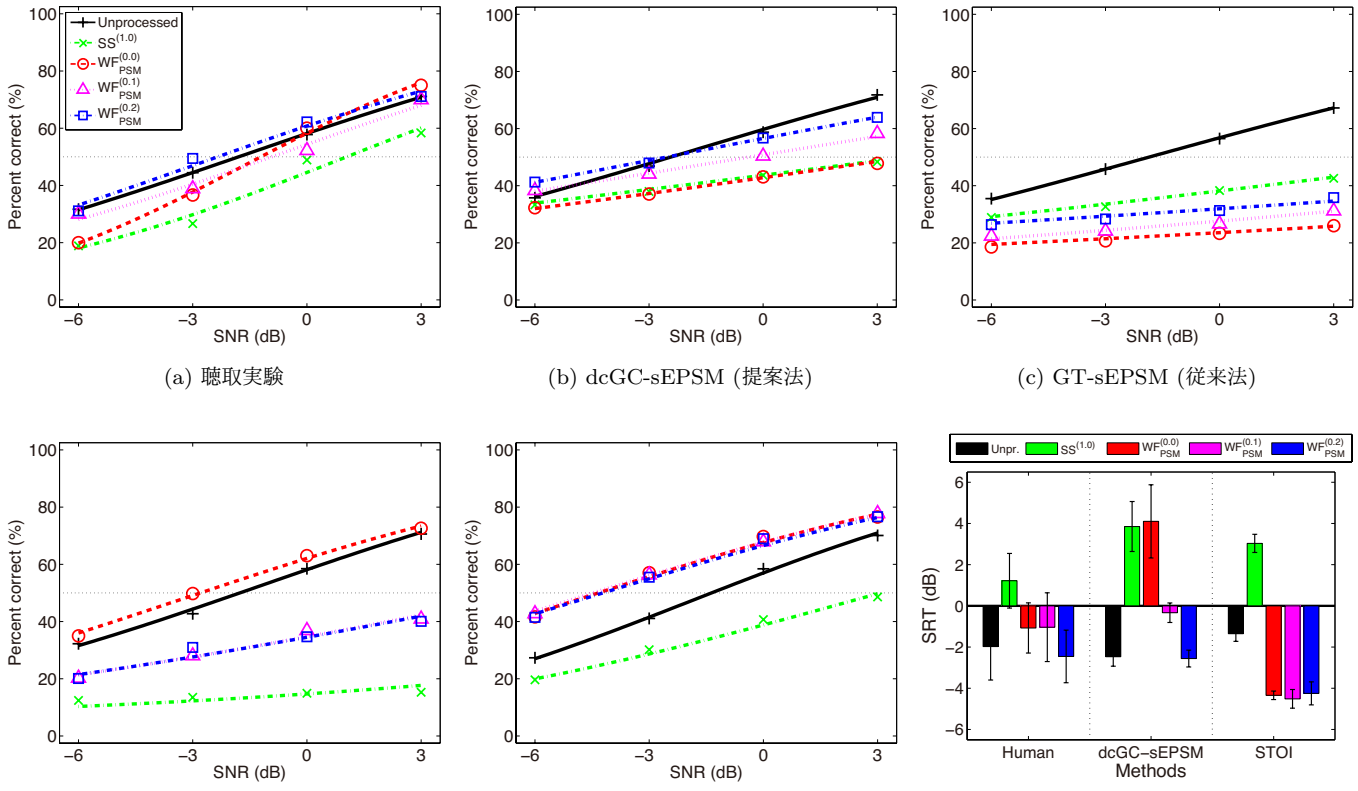
$$P_{correct}^{(STOI)} = \frac{100}{1 + e^{(-7.42\mathrm{STOI}+5.35)}}. \quad (9)$$

4. 結果

4.1 音声明瞭度曲線の分布

図 2 に, 聴取実験から得られた音声明瞭度曲線の分布 (a) と客観的評価指標 (dcGC-sEPSM (b), GT-sEPSM (c), three-level CSII (d), STOI (e)) で予測された音声明瞭度曲線の分布を示す．図中の横軸は，Unprocessed (雑音抑圧処理前の雑音重畳音声) における SNR を表している．聴取実験および各手法の結果は，それぞれ 4 種類の雑音抑圧処理 (スペクトル減算法 $\mathrm{SS}^{(1.0)}$; ウィナーフィルタ型の雑音抑圧法 $\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$, $\mathrm{WF}_{\mathrm{PSM}}^{(0.1)}$, $\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$) に Unprocessed を加えた 5 つの曲線から構成される．図 2 (a) 中のプロットは被験者 9 人分の平均値，図 2 (b)〜(e) 中のプロットは聴取実験に使用した全データごとに算出された音声明瞭度の平均値である．音声明瞭度曲線を得るために，ブートストラップ法 [20, 21] を用いた累積ガウス関数のフィッティングを行った．

聴取実験の結果 (図 (a)) では，$\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$ の音声明瞭度曲線が Unprocessed よりも高い値を示した．対照的に，$\mathrm{WF}_{\mathrm{PSM}}^{(0.1)}$ や $\mathrm{SS}^{(1.0)}$ における音声明瞭度曲線は Unprocessed よりも低い値を示した．$\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$ における音声明瞭度曲線は，SNR が高いときは Unprocessed よりも高く，SNR が低いときは Unprocessed よりも低い値を示した．これらの結果から，聴取実験による知覚的な評価において，$\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$ の雑音抑圧処理が雑音重畳音声の音声明瞭度を改善できることが示唆された．次の節以降では，各客観的評価指標の結果を聴取実験の結果と比較する．



(a) 聴取実験　(b) dcGC-sEPSM (提案法)　(c) GT-sEPSM (従来法)

(d) Three-level CSII　(e) STOI　(f) 音声知覚閾値 (SRT)

図 2: 聴取実験から得られた音声明瞭度 (a), モデルから予測された音声明瞭度 (b)〜(e) と語音認識閾値 (speech reception threshold; SRT) による比較 (f).

提案法である dcGC-sEPSM による音声明瞭度の予測結果 (図 2 (b)) は，全体的に聴取実験の結果 (図 2 (a)) に近い結果となった．全ての雑音抑圧処理に対する音声明瞭度曲線の順序は，$\mathrm{WF}_{\mathrm{PSM}}^{(0.2)} > \mathrm{WF}_{\mathrm{PSM}}^{(0.1)} > \mathrm{SS}^{(1.0)} \simeq \mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$ となり，ほぼ平行の位置関係を示した．Unproessed の音声明瞭度曲線は，SNR が 0 dB 以上のとき他の雑音抑圧処理の条件に対しても高く，SNR が 0 dB 以下のとき $\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$ や $\mathrm{WF}_{\mathrm{PSM}}^{(0.1)}$ よりも低い．これらの結果は聴取実験の結果とやや異なるが，$\mathrm{WF}_{\mathrm{PSM}}$ の上下関係は聴取実験の結果と等しくなった．しかし，$\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$ と $\mathrm{SS}^{(1.0)}$ の音声明瞭度曲線がほぼ同じ位置にあるなど，改善すべき点が見られた．

一方で，GT-sEPSM による音声明瞭度の予測結果 (図 2 (c)) は，聴取実験の結果と大きく異なる結果となった．音声明瞭度曲線の順序は，Unprocessed $\gg \mathrm{SS}^{(1.0)} > \mathrm{WF}_{\mathrm{PSM}}^{(0.2)} > \mathrm{WF}_{\mathrm{PSM}}^{(0.1)} > \mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$ となった．GT-sEPSM が予測した音声明瞭度曲線は，$\mathrm{SS}^{(1.0)}$ と $\mathrm{WF}_{\mathrm{PSM}}$ 間の上下関係が聴取実験の結果と反対である．この傾向は，GT-sEPSM のパラメータを変更しても変えることはできない．

Three-level CSII による音声明瞭度の予測結果 (図 2 (c)) も，聴取実験の結果と大きく異なる結果となった．音声明瞭度曲線の順序は，$\mathrm{WF}_{\mathrm{PSM}}^{(0.0)} >$ Unprocessed $\gg \mathrm{WF}_{\mathrm{PSM}}^{(0.1)} \simeq \mathrm{WF}_{\mathrm{PSM}}^{(0.2)} \gg \mathrm{SS}^{(1.0)}$ となった．3 つの条件の $\mathrm{WF}_{\mathrm{PSM}}$ の上下関係が異なることから，three-level CSII は聴取実験の結果に適合できないと考えられる．

STOI による音声明瞭度の予測結果 (図 2 (e)) は，少なくとも GT-sEPSM や three-level CSII よりも聴取実験の結果に近い結果となった．音声明瞭度曲線の順序は，$\mathrm{WF}_{\mathrm{PSM}}^{(0.1)} \simeq \mathrm{WF}_{\mathrm{PSM}}^{(0.0)} \simeq \mathrm{WF}_{\mathrm{PSM}}^{(0.2)} >$ Unporcessed $> \mathrm{SS}^{(1.0)}$ となった．しかし，聴取実験の結果や提案法の予測結果で見られた $\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$, $\mathrm{WF}_{\mathrm{PSM}}^{(1.0)}$, $\mathrm{WF}_{\mathrm{PSM}}^{(2.0)}$ 間の違いが，STOI ではほとんど現れなかった．このことから，STOI ではスペクトル減算法 (SS) での評価は行えるが，ウィナーフィルタ型雑音抑圧法 ($\mathrm{WF}_{\mathrm{PSM}}$) のパラメータの違いを区別する精度が無いことが考えられる．

4.2 語音認識閾値 (SRT)

図 2 (a)〜(e) において，音声明瞭度曲線が 50%になるときの SNR の値を語音認識閾値 (speech reception threthold; SRT) として計算を行った．全般的に音声明瞭度曲線が高い場合，この SRT は小さい値となる．図 2 (f) に聴取実験 (Human)，客観的評価指標 (dcGC-sEPSM, STOI) の結果からそれぞれ計算された SRT を示す．図中の縦棒と誤差範囲は，聴取実験の結果では実験参加者 9 人分の平均値と標準偏差，客観的評価指標では評価に使用した 9 つのデー

タセット間の平均値と標準偏差を表す．なお，GT-sEPSM (図2(c)) と three-level CSII (図2(d)) では，それぞれの予測結果から計算された SRT が 6 dB 以上になった．これは聴取実験の結果と大きく異なり，この図での比較の意味が無いので掲載していない．

提案法である dcGC-sEPSM の予測結果から得られた SRT は，$\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$ の条件を除いて，聴取実験の結果から得られた SRT と傾向が似ている．一方で，STOI の予測結果から得られた SRT は，$\mathrm{WF}_{\mathrm{PSM}}$ の条件が全体的に低い値を示している上に，パラメータごとの違いが現れていない．以上の結果から，STOI が非線形な雑音抑圧処理手法である $\mathrm{WF}_{\mathrm{PSM}}$ の評価には適していないということがわかる．

dcGC-sEPSM において，$\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$ の SRT の値が過剰に高く（音声明瞭度が過剰に低く）示される原因としては，式 3 における $\mathrm{SNR}_{\mathrm{env}}$ の計算に用いられる $P_{env,\mathrm{N}}$ の曖昧な定義が考えられる．現段階の雑音の残留成分の定義では，過度の処理である $\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$ によって音声状の振幅包絡情報が付与されるため，過剰に低い $\mathrm{SNR}_{\mathrm{env}}$ が計算されている．そのため，今後の研究では雑音の残留成分についてより厳密な定義が必要となる．

5. まとめ

動的圧縮型ガンマチャープフィルタバンクを用いて sEPSM を拡張した音声明瞭度予測法（dcGC-sEPSM）を提案した．スペクトル減算法とウィナーフィルタ型の雑音抑圧法によって強調された音声の明瞭度を聴取実験で測定し，客観的評価指標による予測結果との比較を行った．結果として，dcGC-sEPSM は既存手法 (GT-sEPSM, three-level CSII, STOI) よりも聴取実験の結果に近い予測結果を示した．

謝辞 本研究は，科研費基盤 B 25280063 の支援を一部受けた．

参考文献

[1] ANSI: Methods for Calculation of the Speech Intelligibility Index, ANSI S3.5 (1997).
[2] Steeneken, H. J. M. and Houtgast, T.: A physical method for measuring speech-transmission quality, *J. Acoust. Soc. Am.*, Vol. 67, No. 1, pp. 318–326 (1980).
[3] Rhebergen, K. S. and Versfeld, N. J.: A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners, *J. Acoust. Soc. Am.*, Vol. 117, No. 4, pp. 2181–2192 (2005).
[4] Goldsworthy, R. L. and Greenberg, J. E.: Analysis of speech-based Speech Transmission Index methods with implications for nonlinear operations., *J. Acoust. Soc. Am.*, Vol. 116, No. 6, pp. 3679–3689 (2004).
[5] Kates, J. M. and Arehart, K. H.: Coherence and the speech intelligibility index., *The Journal of the Acoustical Society of America*, Vol. 117, No. 4 Pt 1, pp. 2224–2237 (2005).
[6] Taal, C. H., Hendriks, R. C., Heusdens, R. and Jensen, J.: An algorithm for intelligibility prediction of time-frequency weighted noisy speech, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 19, No. 7, pp. 2125–2136 (2011).
[7] Jørgensen, S. and Dau, T.: Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing, *J. Acoust. Soc. Am.*, Vol. 130, No. 3, pp. 1475–1487 (2011).
[8] 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の改良，聴覚研究会資料，Vol. 46, No. 1, 日本音響学会，pp. 35–40 (2016).
[9] Patterson, R. D., Allerhand, M. H. and Giguère, C.: Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform, *J. Acoust. Soc. Am.*, Vol. 98, No. 4, pp. 1890–1894 (1995).
[10] Irino, T. and Patterson, R. D.: A Dynamic Compressive Gammachirp Auditory Filterbank., *IEEE Trans. Audio. Speech. Lang. Processing*, Vol. 14, No. 6, pp. 2222–2232 (2006).
[11] Mickes, L., Wixted, J. T. and Wais, P. E.: A direct test of the unequal-variance signal detection model of recognition memory., *Psychon. Bull. Rev.*, Vol. 14, No. 5, pp. 858–65 (2007).
[12] Green, D. M. and Birdsall, T. G.: The effect of vocabulary size, *Signal Detection and Recognition by Human Observers*, Wiley, New York, pp. 609–619 (1964).
[13] Berouti, M., Schwartz, R. and Makhoul, J.: Enhancement of speech corrupted by acoustic noise, *ICASSP '79. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Vol. 4, Institute of Electrical and Electronics Engineers, pp. 208–211 (1979).
[14] Fujimoto, M., Watanabe, S. and Nakatani, T.: Noise suppression with unsupervised joint speaker adaptation and noise mixture model estimation, *2012 IEEE Int. Conf. Acoust. Speech Signal Process.*, IEEE, pp. 4713–4716 (2012).
[15] Moreno, P. J., Raj, B. and Stern, R. M.: A vector Taylor series approach for environment-independent speech recognition, *Acoust. Speech, Signal Process. 1996. ICASSP-96. Conf. Proceedings., 1996 IEEE Int. Conf.*, Vol. 2, IEEE, pp. 733–736 (1996).
[16] Amano, S., Kondo, T., Suzuki, Y. and Sakamoto, S.: Familiarity-controlled word lists 2007 (FW07), The Speech Resources Consortium, National Institute of Informatics (2007).
[17] Nelder, J. A. and Mead, R.: A Simplex Method for Function Minimization, *The Computer Journal*, Vol. 7, No. 4, pp. 308–313 (1965).
[18] Amano, S. and Kondo, T.: Estimation of mental lexicon size with word familiarity database, *Int. Conf. Spok. Lang. Process.*, pp. 2119–2122 (1998).
[19] Moore, B. C. J.: Suggested formulae for calculating auditory-filter bandwidths and excitation patterns, *The Journal of the Acoustical Society of America*, Vol. 74, No. 3, p. 750 (1983).
[20] Wichmann, F. A. and Hill, N. J.: The psychometric function: I. Fitting, sampling, and goodness of fit, *Perception & psychophysics*, Vol. 63, No. 8, pp. 1293–1313 (2001).
[21] Wichmann, F. A. and Hill, N. J.: The psychometric function: II. Bootstrap-based confidence intervals and sampling, *Perception & psychophysics*, Vol. 63, No. 8, pp. 1314–1329 (2001).