音響信号符号化のための周波数軸を伸縮するスペクトル包絡表現*

☆杉浦亮介[1], 鎌本優[2], 原田登[2], 亀岡弘和[2], 守谷健弘[2] ([1]東大院 情報理工, [2]NTT CS 研)

1 はじめに

音声音響符号化は携帯電話やインターネット電話 (VoIP) 等の音声通信において不可欠な技術であり, 限られたリソース下においては, 符号化方式が通信の質を大きく左右させる. 本稿では, その中でも主に携帯電話で使われるような, 実時間で低遅延かつ非可逆な圧縮を行う符号化について取り扱う.

近年携帯電話の符号化では扱う音の広帯域化が進められ, 今まで不得意であった音声以外の音響信号への対応が進められている. 音声と音響に対応した符号化方式の標準規格としては, 3GPP Extended Adaptive Multi-Rate Wideband (AMR-WB+) や MPEG-D Unified Speech and Audio Coding (USAC) [1, 2] が知られている. これらはいずれも入力が音声なら時間領域, 音声以外なら周波数領域での符号化を行う. 本研究では後者の周波数領域での符号化である Transform Coded eXitation (TCX) に焦点を当て, 低ビットレート, 低遅延の条件下での音質向上を目指す.

TCX では信号スペクトルの包絡情報が符号化の効率に大きく影響を与える. そこで本稿では, 周波数軸の伸縮を用いたスペクトル包絡の効率的な表現法, 及びその際の聴覚的な重み付けについての提案を行う.

2 周波数領域での符号化 (TCX)

TCX の手法の中で現在最も有力視されているのは, 修正離散コサイン変換 (MDCT) を用いるもの [3] であり, これは USAC でも採用されている. 処理の概略を図 1 で示す. TCX では大きく二つの情報を量子化して送る. 一つはスペクトルの包絡を表す線形予測係数, もう一つはスペクトルをその包絡で割って得られる残差スペクトルである. 線形予測係数は, 量子化や内挿に対して頑健な線スペクトル対 (LSP) に変換してベクトル量子化し, 残差は周波数ビン毎にスカラー量子化後, エントロピー符号化し圧縮する.

残差を計算する際には聴覚的な重み付けが行われる. 線形予測係数を $\{a_n\}$ とすると, スペクトル包絡は

$$H(k) = 1/|1+\sum_n a_n e^{-j\frac{2\pi k}{N}n}|, (0 \le k \le N-1) \quad (1)$$

と表され, 残差を求める際には, この包絡を重み付けにより平滑化した

$$\tilde{H}(k) = 1/|1+\sum_n a_n \gamma^n e^{-j\frac{2\pi k}{N}n}|, (0 < \gamma < 1) \quad (2)$$

で元のスペクトルを除算する. この残差をスカラー量子化すると, スペクトルの量子化歪みはおよそ $\tilde{H}(k)/H(k)$ に比例し, スペクトルのピークでは歪が小さく, その周辺の周波数では歪みが大きくなり, 聴感的に最適な歪みの分配がなされる. γ は実験的に 0.92 を使うと聴感に適うとされている.

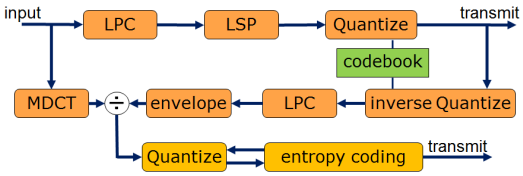


Fig. 1 TCX のフローチャート

3 提案手法

3.1 周波数軸伸縮を用いたスペクトル包絡表現

上述のような聴感的に効率のよい量子化誤差の分配を行うためには, スペクトル包絡が正しく表現されている必要がある.

一般的に, 線形予測分析で得られる包絡はおおよそ (信号の点数)/(線形予測次数) の解像度で元のスペクトル表現し, この解像度は周波数において均一である. しかし, 実際の音響信号の多くは低周波に聴感上重要な構造を有する. また, 人の聴覚の周波数分解能は対数的であるとされている. したがって, スペクトル包絡の解像度を伸縮することにより, スペクトル包絡の表現効率が上がる可能性があると考えられる.

周波数毎に解像度の違う包絡のモデルとしては, 今井らの Mel Log Spectrum Approximation (MLSA) フィルタ [4] が代表的である. このフィルタは, 包絡が対数周波数領域で均一な分解能になるような近似がなされている. しかし, このモデルでは係数の推定のためにフレームごとの反復計算が必要となり, 包絡の計算にも演算量がかかることから, 実時間の周波数領域符号化には向かないと考えられる.

そこで本稿では, 周波数軸の伸縮と逆伸縮を近似する行列を作成し, それらを用いて解像度が伸縮された包絡を形成するモデルを提案する. 提案手法の概要を図 2 に示す. まず, 行列演算により擬似的にパワースペクトルを対数周波数軸に伸縮する. そしてこの伸縮されたスペクトルをフーリエ変換したものを信号の自己相関関数として, 線形予測分析により線形予測係数を得る. その係数に対して式 (1) を適用すると対数周波数軸上での形状を表す包絡が求まるので, 逆伸縮の行列演算を行い, 元の周波数軸に戻す. この手法は線形予測分析により推定を行うので, 係数は陽に求まり, 必ず安定であることが保証できる.

---

*"Frequency warped spectral envelope representation for audio coding." by Ryosuke Sugiura[1], Yutaka Kamamoto[2], Noboru Harada[2], Hirokazu Kameoka[2] and Takehiro Moriya[2] ([1]The University of Tokyo, [2]NTT Communication Science Laboratories).

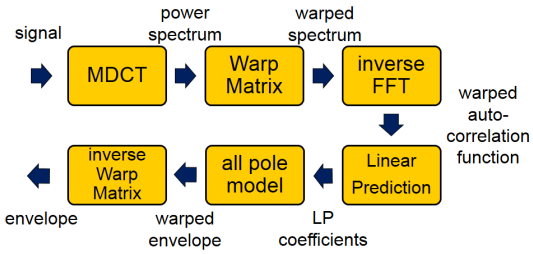


Fig. 2　提案手法のフローチャート

3.2　周波数伸縮の行列近似

そもそも, 離散信号における周波数伸縮は不可逆的な操作であり, 単純に sinc 補間を使用した場合, 意図せぬ変形がおこる可能性がある. そこで, 周波数伸縮とその逆伸縮を近似する行列を予め学習によって求めておくことを考える. 線形予測分析から求まる包絡は, その自乗値と元のパワースペクトルとの板倉斎藤距離が最小となるものである [5] ことから, 本手法の学習でも同様に目的関数として板倉斎藤距離を用いる. 周波数伸縮行列 W に関しては,

$$\sum_{i,j}\left(\frac{Y_{ij}}{\sum_k W_{ik}X_{kj}} - \log\frac{Y_{ij}}{\sum_k W_{ik}X_{kj}} - 1\right) \tag{3}$$

を目的関数とし, 行列 W について最小化する. ただし, X は学習データのパワースペクトル, Y は sinc 補間によって周波数伸縮を行ったパワースペクトルであり, 目的関数は WX の Y からの板倉斎藤距離を示す. また, パワースペクトルを伸縮する行列であることから, W の要素は非負値制約で求める必要がある.

この最適化問題は解析的に解けないので, 補助関数法 [6] による反復計算を用いて最適解を求める. まず, この目的関数の W に関する項のみについて考え,

$$\sum_{i,j}\frac{Y_{ij}}{\sum_k W_{ik}X_{kj}} + \sum_{i,j}\log\left(\sum_k W_{ik}X_{kj}\right) \equiv L(W) \tag{4}$$

を最小化することを考える. 関数 $f(x) = \frac{1}{x}$ は $x > 0$ において凸であるので Jensen の不等式

$$\begin{aligned}\frac{1}{\sum_k W_{ik}X_{kj}} &= \frac{1}{\sum_k \lambda_{ijk}(W_{ik}X_{kj}/\lambda_{ijk})} \\ &\le \sum_k \frac{\lambda_{ijk}}{W_{ik}X_{kj}/\lambda_{ijk}} \quad \left(\lambda_{ijk} \ge 0,\ \sum_k \lambda_{ijk} = 1\right)\end{aligned} \tag{5}$$

が成り立つ. また, 対数関数の凸性から

$$\log\left(\sum_k W_{ik}X_{kj}\right) \le \log\phi_{ij} + \frac{\sum_k W_{ik}X_{kj}}{\phi_{ij}}, (\phi_{ij} > 0) \tag{6}$$

が導かれる. この二式を用いて, 上記の目的関数の W に関する項の上界を次の補助関数により定める.

$$\begin{aligned}L(W) \quad \le \quad & \sum_{i,j} Y_{ij}\sum_k \frac{\lambda_{ijk}^2}{W_{ik}X_{kj}} \\ & + \sum_{i,j}\left(\log\phi_{ij} + \frac{\sum_k W_{ik}X_{kj}}{\phi_{ij}}\right) \equiv G(W)\end{aligned} \tag{7}$$

ただし等号成立は

$$\lambda_{ijk} = \frac{W_{ik}X_{kj}}{\sum_k W_{ik}X_{kj}},\ \ \phi_{ij} = \sum_k W_{ik}X_{kj} \tag{8}$$

の時である. 式 (7) は凸であるので, λ, ϕ を固定すると上記の補助関数 $G(W)$ は W に関する停留点を求めることにより最小化できる. よって,

$$\begin{aligned}&\frac{\partial}{\partial W_{mn}}G(W)|_{W=\tilde{W}} \\ =\ & \sum_j Y_{mj}\lambda_{mjn}^2/X_{nj}\left(-\frac{1}{\tilde{W}_{mn}^2}\right) + \sum_j\frac{X_{nj}}{\phi_{mj}} = 0\end{aligned}$$

$$\iff \tilde{W}_{mn} = \sqrt{\frac{\sum_j Y_{mj}\lambda_{mjn}^2/X_{nj}}{\sum_j X_{nj}/\phi_{mj}}} \tag{9}$$

により, この $\tilde{W}$ が補助関数 $G(W)$ を最小化する. そして $\tilde{W}$ と式 (8) により λ, ϕ を更新し, 再び式 (7) を最小化する. これを繰り返すことにより目的関数は単調に減少し, $\tilde{W}$ は局所解に近づく. この反復は次のように更新式にまとめられる.

l 回目の反復で得られる $\tilde{W}$ を $W^{(l)}$ とすると, それにより更新される λ, ϕ はそれぞれ

$$\lambda_{ijk} = \frac{W_{ik}^{(l)}X_{kj}}{\sum_k W_{ik}^{(l)}X_{kj}},\ \ \phi_{ij} = \sum_k W_{ik}^{(l)}X_{kj} \tag{10}$$

となり, これと式 (9) により補助関数 $G(W)$ を最小化する $W^{(l+1)}$ を求めると,

$$W_{mn}^{(l+1)} = \sqrt{\frac{\sum_j Y_{mj}W_{mn}^{(l)2}X_{nj}/\left(\sum_k W_{mk}^{(l)}X_{kj}\right)^2}{\sum_j X_{nj}/\sum_k W_{mk}^{(l)}X_{kj}}} \tag{11}$$

が得られ,

$$W_{mn}^{(l+1)} = W_{mn}^{(l)}\sqrt{\frac{\sum_j Y_{mk}X_{nj}/\hat{Y}_{mj}^2}{\sum_j X_{nj}/\hat{Y}_{mj}}},\ \ \hat{Y} = W^{(l)}X \tag{12}$$

という更新式が求まる. 逆変換の行列 U も同様に, UWX の X からの板倉斎藤距離最小化を考えることで学習の更新式が求められる.

上記で得られる更新式は, 正数との積の形で表される. したがって, 行列 W, U の要素は初期値が正数であれば必ず正数となり, 零であれば更新後も零である. このことから, 初期値の要素に零を使うことで周波数伸縮演算のタップ数を制限し, 演算量を調節することができる. また, 学習データ Y を変えることで周波数伸縮を自由に設計することもできる.

以下では W, U として, 各行の非零要素数を高々7つまでとして学習した対数伸縮行列を使用する.

3.3　伸縮された周波数軸上での聴覚重み付け

提案手法で得られた包絡を実際の TCX で使用する場合, 先述のとおり式 (2) の聴覚重み付けによる平滑化を行う必要がある. しかし, 伸縮を行った場合はこの平滑化に注意しなければならない. 提案手法によって得られた線形予測係数を式 (2) に適用すると, 対数周波数軸上のスペクトル包絡を, 線形周波数軸上の包

絡として平滑化したものとなる．したがって本来の周波数軸での平滑化とは結果が異なる．実験的に有効性が知られている $\tilde{H}(k)/H(k)$ に従う量子化歪みの分配を行うためには，その周波数軸にあった方法で $\tilde{H}(k)$ を求める必要がある．そこで，周波数ビン毎に次のような補正を γ にかける．

対数周波数軸と線形周波数軸において，ナイキスト周波数までの区間にそれぞれ等間隔に N 点の離散周波数があると考える．そして，式 (2) の k を対数周波数軸上での離散周波数のインデックスとし，$f(k)$ を，k 番目の対数周波数が線形周波数軸において何番目に相当するかを表すものとする．ただし，$f(0)=0$ とする．伸縮された包絡の平滑化の際，提案手法で得られたパラメタ系列 $\{a_n\}$ を用いて，

$$\tilde{H}(0) = 1/|1+\sum_n a_n\gamma^n|$$

$$\tilde{H}(k) = 1/|1+\sum_n a_n\gamma^{\frac{f(k)}{k}n}e^{-j\frac{2\pi k}{N}n}|, (1\le k\le N-1) \quad (13)$$

のように γ に周波数軸伸縮の度合いに応じた補正をかけることで，解像度が均一な場合の包絡の平滑化を近似する．

また，上記の $f(k)$ は逆伸縮行列 U を使うことで，

$$\begin{pmatrix} f(0) \\ \vdots \\ f(N-1) \end{pmatrix} \simeq U \begin{pmatrix} 0 \\ \vdots \\ N-1 \end{pmatrix} \quad (14)$$

と近似でき，この場合でも同様の効果が得られる．

4 実験と結果

4.1 スペクトル包絡の比較

前節の提案手法を評価するため，スペクトル包絡の比較を行った．音響信号の各フレームに対して，提案手法による周波数伸縮，sinc 補間による周波数伸縮，及び MLSA フィルタをそれぞれ用いて 16 次の包絡を求め，パワースペクトルとの板倉斎藤距離を計算した．

図 3 はある 1 フレームで線形予測分析と提案手法による包絡を並べたものである．低域において提案手法の包絡の解像度が向上していることが確認できる．

また，定量的な評価として，各手法による包絡が通常の線形予測分析による包絡と比べ距離がどれだけ近づくかを比較した結果を図 4 に示す．各手法はいずれも通常の線形予測分析と比べ，人の聴感上重要である低域での精度が上がり，その分高域での精度が下がっている．学習した行列と sinc 補間との比較 (a) では，いずれも低域での精度向上は同等であるが，学習した方では高域において包絡の補間が行われ，精度低下が抑えられていることがわかる．そして MLSA フィルタによる包絡との比較 (b) では，提案手法は MLSA フィルタとほぼ同等な性能を示した．

4.2 聴覚重み付け比較

続いて，聴覚重み付けによる包絡の平滑化に関しても比較を行った．通常，線形予測分析で得られた包絡

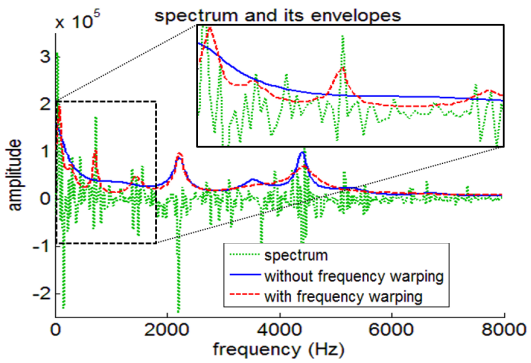


Fig. 3　包絡の比較．緑点線が MDCT による実数スペクトル，赤破線が提案手法，青実線が線形予測分析による包絡．

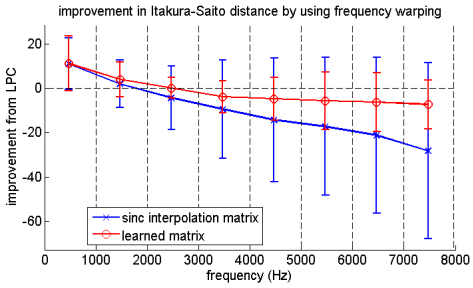
を式 (2) によって重み付けした際には，包絡のピークが急峻なほど大きく平滑化され，なだらかなものはあまり変化しない．しかし，図 5 のように，提案手法で得られた包絡に同様の操作を行った場合，低域において通常よりもピークの形が大きく残った．これは，式 (2) の操作が対数周波数軸上で行われてしまうことに起因する．そこで，式 (13) を用いることで，図 6 のように軸の伸縮に合わせて平滑化された包絡が得られた．周波数毎に γ に補正をかけることにより，線形周波数軸上でのピークの急峻さにあわせて平滑化されることが確認できた．

4.3 主観評価実験

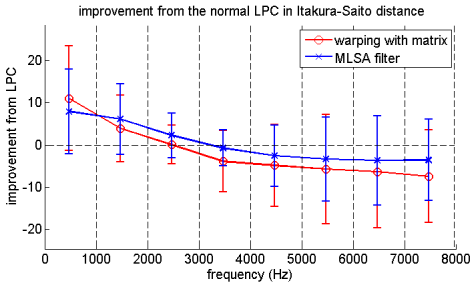
最後に TCX をベースとしたコーデックを作成し，ITU-R BS.1534-1 MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) による音質の主観評価実験を行った．作成した TCX コーデックと AMR-WB+はフレーム長が 16 kHz サンプリングでそれぞれ 320 点 (20 ms) と 1024 点 (64 ms) であり，アルゴリズム遅延がそれぞれ 40 ms, 64 ms〜72 ms 発生する．楽曲は RWC 音楽データベースの異なるジャンルの 6 曲から 10 秒を切り取り，16 kHz にダウンサンプルした後 16 kbps で圧縮した．

7 名の被験者にそれぞれ楽曲について参照音と TCX の周波数伸縮有り・無し，及び参照方式として AMR-WB+によってそれぞれ圧縮した音，そしてアンカーとして 3.5 kHz に帯域制限した音を提示し，どれだけ参照音に近い音であるか 100 点満点で評価を行った．

図 7(a) は各圧縮方式についての評価結果である．作成したコーデックと AMR-WB+は，それぞれ曲調による得手不得手があるがほぼ同等の評価が得られた．また，周波数伸縮をしたことによる点数の増加を図 7(b) で示す．曲調によって効果の程度は異なるが，6 曲全てにおいて伸縮有りは無しに比べて同等以上の評価が得られ，3 曲においては有意水準 5 % で有意差が見られた．このことから，周波数伸縮が音質の向上に寄与していることが言える．



(a) sinc 補間との精度比較. 赤丸が学習した行列, 青クロスが sinc 補間によるもの.



(b) MLSA フィルタとの精度比較. 赤丸が提案手法, 青クロスが MLSA フィルタ. MLSA フィルタは Speech Signal Processing Toolkit[7] のものを使用.

Fig. 4　各帯域での包絡の線形予測分析からの改善量 (板倉斎藤距離基準) の平均と標準偏差. 縦軸が 0 より大きい所は, 線形予測分析の包絡よりもスペクトルからの距離が近いことを意味する. RWC 研究用音楽データベースのポピュラー音楽及びクラシック計 15 曲からそれぞれランダムに 30 秒を切り取り使用. 16 kHz サンプル, 1 フレームあたり 320 点のスペクトル, 包絡の次数は 16 次.

5　おわりに

本稿では, 周波数伸縮とそれに対する逆伸縮を近似する行列を用いた包絡モデルを提案した. 提案手法を用いることで, 解像度の伸縮された包絡を形成することができ, 対数軸に伸縮した場合は MLSA フィルタとほぼ同等な精度となることを確かめた. また, 周波数伸縮を使用する際には, 聴覚重み付けの仕方に修正を加える必要があることを指摘した. そして, 音質の主観評価実験により, 周波数伸縮の効果を統計的に示した.

包絡の解像度伸縮がどのような場合において有効であるか, また, 対数周波数軸以外で効果のある周波数伸縮があるかについての調査が今後の課題である.

参考文献

[1] 3GPP TS 26.290, 3GPP, 2012.
[2] M. Neuendorf, et al., AES 132nd Convention, Budapest, 2012.
[3] G. Fuchs, et al., EUSIPCO, pp.1264-1268, 2009.
[4] 今井, et al., 電子通信学会論文誌 '83/2, Vol.J66-A, No.2, pp.122-129, 1983.
[5] 板倉, 博士論文, 名大院工学研究科, 1972.
[6] 亀岡, et al., 情処研報, vol.66, pp.77-84, 2006.
[7] http://sp-tk.sourceforge.net/ ('13 年 1 月現在).

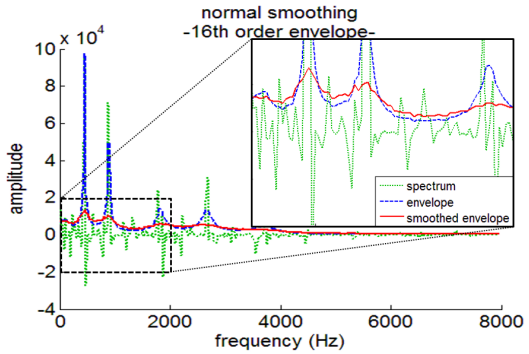


Fig. 5　提案手法の包絡を式 (2) で平滑化した場合. 青破線が平滑化前, 赤実線が平滑化後, 次数は 16.

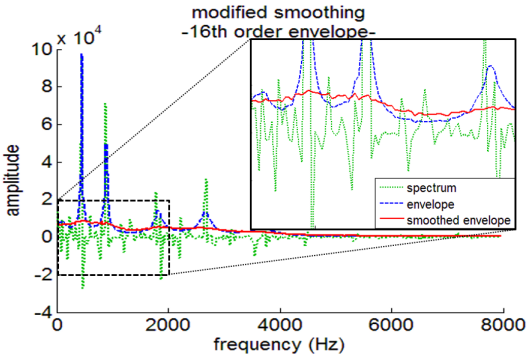
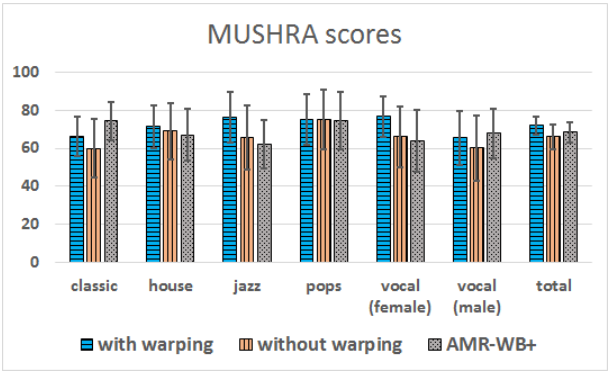
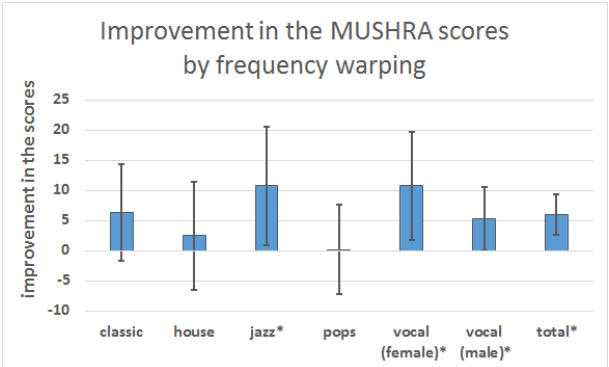


Fig. 6　提案手法の包絡を式 (13) で平滑化した場合. 青破線が平滑化前, 赤実線が平滑化後, 次数は 16.



(a) MUSHRA スコア. 左から順に周波数伸縮有り, 無し, AMR-WB+.



(b) 周波数伸縮を用いたことによる点数の増加量.

Fig. 7　主観評価実験の結果. エラーバーは 95 %信頼区間, *は 5 %有意水準で有意差があったことを示す.