音響符号化のための周波数解像度の伸縮されたスペクトル包絡の表現法

杉浦 亮介[1,a)]　鎌本 優[2,b)]　原田 登[2,c)]　亀岡 弘和[2,d)]　守谷 健弘[2,e)]

概要：本稿では, 非負値疎行列を用いた周波数軸伸縮により周波数解像度を伸縮させることで, 効率よく周波数スペクトル包絡を表現するためのモデルを提案する. このモデルは予め伸縮のための行列を設計しておくことにより, 包絡の解像度を任意に伸縮させることができ, 周波数領域での音声音響符号化において入力信号の情報の偏りに合わせて包絡を表現し, 符号化器の品質を向上させることが期待できる. ここではその一端としてメル対数的な解像度の伸縮を実際の符号化器に取り入れ, 低ビットレート・低遅延の符号化器において品質が改善されることを主観評価実験の結果により示す.

キーワード：音響符号化, 信号処理, 周波数伸縮, 非負値疎行列, TCX

1. はじめに

音声符号化の技術は長年に渡って多くの研究がなされ, 音声通信の根底を支えてきた. これまでの音声通信用の符号化方式は比較的狭い帯域を扱い, 主に音声信号に特化して設計されてきた. しかし, 携帯電話やインターネット電話 (VoIP) 等による音声通信が当たり前になった昨今においては, より快適なコミュニケーションの実現のため, 広帯域な信号を対象とし, 音声以外の信号も高品質に圧縮できる方式が求められている.

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といったように, 符号化を行う領域を使い分けている. 本研究は, 主に携帯電話での使用を想定し, AMR-WB+や USAC よりも低遅延な条件下での高品質な音声音響統合符号化方式を最終目標とする. 本稿ではその一端として, 周波数領域での符号化である Transform Coded eXitation (TCX) に焦点を当て, 品質改善のための提案を行う.

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. 本稿では, 非負値疎行列を用い, 低演算量で包絡の解像度を伸縮することにより, TCX にとって効率のよいスペクトル包絡の表現を行うことを考える.

[1] 東京大学大学院情報理工学系研究科
[2] NTT コミュニケーション科学基礎研究所
[a)] sugiura@hil.t.u-tokyo.ac.jp
[b)] kamamoto.yutaka@lab.ntt.co.jp
[c)] harada.noboru@lab.ntt.co.jp
[d)] kameoka.hirokazu@lab.ntt.co.jp
[e)] moriya.takehiro@lab.ntt.co.jp

2. 周波数領域符号化におけるスペクトル包絡

本研究は, USAC でも採用されている修正離散コサイン変換 (MDCT) を用いた TCX [3] をベースに議論を行う. TCX の処理の概略を図 1 で示す. この方式ではスペクトルの包絡を表す線形予測係数, 及び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を量子化し, 符号化する. 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.

ここで, スペクトル包絡の情報は大きく二つの役割を果たす. まず一つ目は残差スペクトルの量子化の際に生じる誤差の分配である. 信号の残差スペクトルを計算する際, 包絡に聴覚的な重み付けを行うことで量子化誤差を調節することができる. 通常, スペクトル包絡は線形予測係数を $\{a_n\}$ を用いて

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

と表されるが, この包絡の代わりに H_k を重み付けにより平滑化した

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

を除算に用いて残差を得る. 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

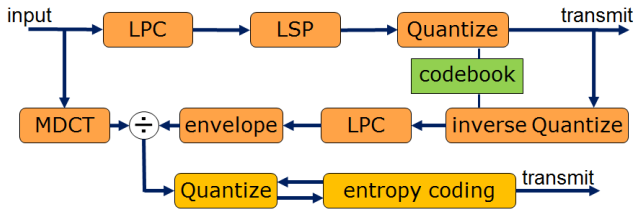


図 1　TCX のフローチャート

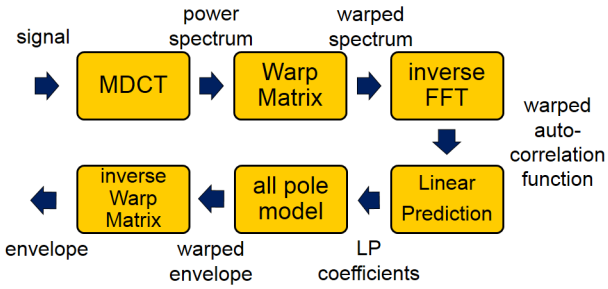


図 2　提案手法のフローチャート

は歪みが大きくなるように歪みの分配がなされる．上記の γ として 0.92 を使うと，おおよそ人間の聴覚におけるマスキング効果を考慮した誤差配分になることが実験的に知られており，様々な符号化方式でこの値が用いられている．

二つ目の役割は残差のエントロピー符号化におけるパラメータ調節である．残差スペクトルの包絡は $H_k/\tilde{H}_k$ で表すことができ，この包絡と残差の値には相関が見られる．例えばエントロピー符号化としてライス符号 [4] を用いる場合，この包絡の対数に比例した値を各ビンのライスパラメータとすれば符号化の効率を上げることができる．

3. 提案手法

3.1 周波数軸伸縮を用いた包絡の解像度伸縮

上記のような誤差分配やパラメータ調節はスペクトル包絡が元のスペクトルの構造を正しく表現していることを前提としたものであり，包絡が適切に表現されていない時には逆効果となり得る．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することから，包絡の次数を上げることにより包絡の精度自体は上昇させることが可能であるが，その分伝送すべき情報が増えてしまい必ずしも符号化効率の改善には寄与しない．そこで本稿では，自然音のエネルギーの偏りに着目し，同次数でより効率よく情報を抽出することを考える．高圧縮の符号化においては多くの場合，量子化の過程で大半の帯域の情報がなくなり，もともとエネルギーの偏っていた帯域の情報が主にエントロピー符号化の対象となる．この事実から，解像度に偏りを持たせた包絡モデルを基にスペクトル包絡の情報を抽出することにより，符号化器の効率が上がり得ると考えられる．

解像度の伸縮された包絡のモデルとしては，今井らの Mel Log Spectrum Approximation (MLSA) フィルタ [5] が代表的である．このモデルは，包絡がメル対数周波数軸で均一な分解能になるような近似がなされており，時間領域符号化への応用も検討されている [6], [7]．ただし，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上，伸縮の自由度も小さい．

より低演算量で自由な解像度の伸縮を行うため，本稿では周波数軸の伸縮と逆伸縮を近似する行列による包絡のモデルを提案する．提案手法の概要を図 2 に示す．まず，予め設計した行列により近似的にパワースペクトルを伸縮し，フーリエ変換したものを信号の擬似的な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．得られた係数に対して式 (1) を適用すると伸縮周波数軸上で一様な解像度を持つ包絡が求まる．したがって，逆伸縮の行列演算を行い元の周波数軸に戻すことで，周波数解像度の伸縮された包絡が得られる．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．

3.2 非負値疎行列による周波数伸縮の近似

上述した周波数伸縮の近似行列の設計の際には，次の三つの要素について考慮する．まず，伸縮操作の演算量を抑えるため，伸縮行列は疎なものでなければならない．また，伸縮後のスペクトルもパワースペクトルでなければならないことから，伸縮行列の要素は全て正とするべきである．そして，伸縮・逆伸縮における無矛盾性を保つ必要がある．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．以上のことを踏まえ，周波数伸縮とその逆伸縮を近似する行列を予め最適に設計し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ことから [8] ，行列の最適化において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，

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

を目的関数とし，行列 W について非負値疎行列という条件下で最小化する．ただし，X はトレーニングデータのパワースペクトル，Y は sinc 補間によって厳密に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．

この最適化問題は解析的に解けないことから，補助関数法 [9] 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，

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

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

$$\frac{1}{\sum_k W_{ik}X_{kj}} = \frac{1}{\sum_k \lambda_{ijk}(W_{ik}X_{kj}/\lambda_{ijk})} \leq \sum_k \frac{\lambda_{ijk}}{W_{ik}X_{kj}/\lambda_{ijk}} \quad \left(\lambda_{ijk}\geq 0,\ \sum_k \lambda_{ijk}=1\right) \tag{5}$$

が成り立つ．また, 対数関数の凹性から

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

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．

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

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,

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

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

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更新式が求められる．

上記で得られる更新式は, 行列の各要素と正数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, 初期値の非零要素数を少なくし全て正数とすることにより, 伸縮行列が非負値疎行列であるという条件下で最適化することができる．また, トレーニングデータ Y の伸縮を変えることで伸縮行列を自由に設計することもできる．

3.3 解像度伸縮包絡に対する聴覚重み付け

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, 提案手法によって得られた線形予測係数を単純に式 (2) に適用すると, 伸縮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り, 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, その伸縮にあわせた方法で包絡を平滑化する必要がある．そこで, 周波数ビン毎に次のような補正を γ にかける．

伸縮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 を伸縮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,

$$\tilde{H}_0 = 1/|1+\sum_n a_n\gamma^n|$$
$$\tilde{H}_k = 1/|1+\sum_n a_n\gamma^{\frac{f(k)}{k}n}e^{-j\frac{2\pi k}{N}n}|,\ (1\leq k\leq N-1) \tag{13}$$

のように γ に伸縮の度合いに応じた補正をかけることで, 通常の包絡の平滑化に相当する操作を近似する．

また, 上記の $f(k)$ は逆伸縮行列 U を使うことで,

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

としても同様の効果が得られる．

4. 実験と結果

4.1 メル対数伸縮における比較

最適化により得られた伸縮の近似行列による解像度伸縮の効果を評価するため, メル対数周波数軸への伸縮を用いた時の包絡の比較を次のように行った．まず, 前節の最適

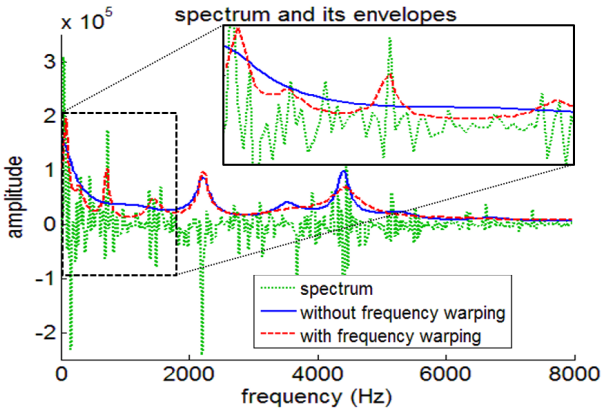


図 3 包絡の比較. 緑点線が MDCT による実数スペクトル, 赤破線が提案手法, 青実線が線形予測分析による包絡.

化により各行の非零要素数が最大 7 つであるようなメル対数伸縮行列とその逆伸縮行列を設計した. そして, 音響信号の各フレームに対して, 最適化した伸縮行列または厳密な sinc 補間による伸縮を用いた二種類の解像度伸縮包絡, 及び MLSA フィルタによる包絡を求め, 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最適化した伸縮行列を用いた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各帯域において精度がどれだけ変化するかを比較した結果を図 4 に示す. 各手法の包絡はいずれも対数的に周波数解像度を伸縮していることから, 通常の線形予測分析と比べ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. 最適化した行列と厳密な sinc 補間との比較では, いずれも低域での精度向上は同等であるが, 最適化した方では sinc 補間よりも高域での伸縮・逆伸縮における無矛盾性が保たれていることがわかる. そして, 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, 前節と同様の解像度伸縮をした包絡にこの単純な平滑化を行った結果, 図 5 のように低域において通常よりもピークの形が大きく残ってしまった. これは式 (2) の操作がメル対数周波数軸上で行われてしまうことに起因する. 一方, 式 (13) を用いた場合, 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.

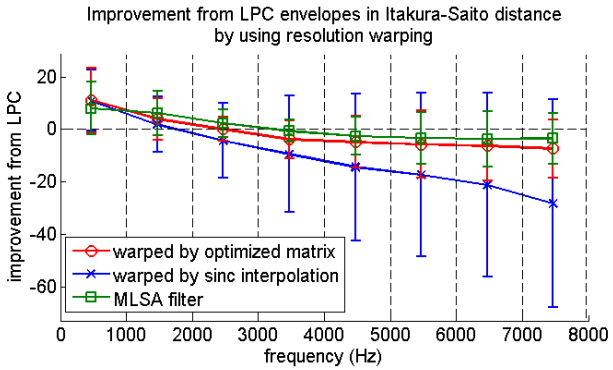


図 4 各帯域での包絡の線形予測分析からの改善量 (板倉斎藤距離基準) の平均と標準偏差. 縦軸が 0 より大きい所は, 線形予測分析の包絡よりもスペクトルからの距離が近いことを意味する. RWC 研究用音楽データベースのポピュラー音楽及びクラシック計 15 曲からそれぞれランダムに 30 秒を切り取り使用. サンプリング周波数 16 kHz, 1 フレームあたり 320 点のスペクトル, 包絡の次数は 16 次. MLSA フィルタは Speech Signal Processing Toolkit [10] のものを使用.

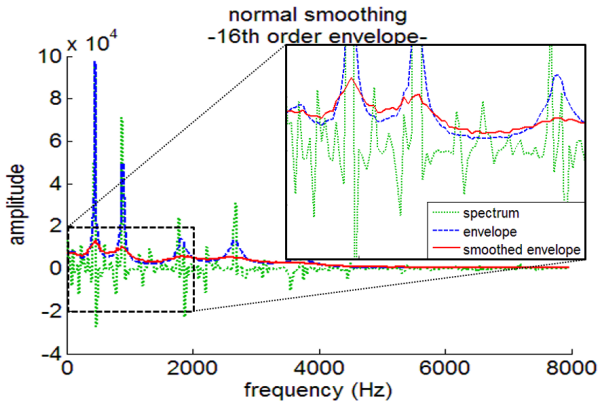


図 5 提案手法の包絡を式 (2) で平滑化した場合. 青破線が平滑化前, 赤実線が平滑化後, 次数は 16.

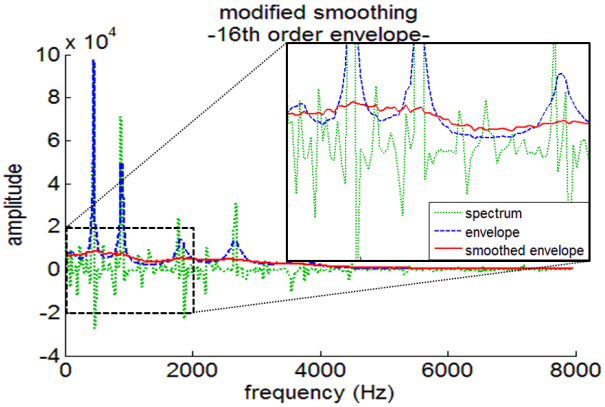
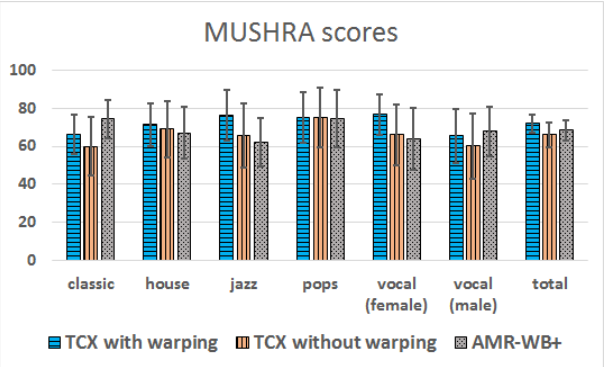


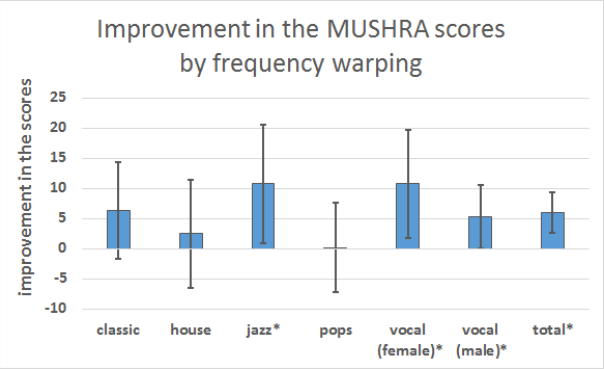
図 6 提案手法の包絡を式 (13) で平滑化した場合. 青破線が平滑化前, 赤実線が平滑化後, 次数は 16.

4.3 音質の主観評価

最後に TCX をベースとした符号化器を作成し, 上の実験で使用したメル対数伸縮行列による解像度伸縮を導入して, ITU-R BS.1534-1 MUltiple Stimuli with Hidden Ref-



(a) MUSHRA スコア. 左から順に包絡の解像度伸縮有りの TCX, 無しの TCX, AMR-WB+.



(b) 周波数伸縮を用いたことによる点数の増加量.

図 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.

erence and Anchor (MUSHRA) による音質の主観評価実験を行った. 作成した TCX はサンプリング周波数 16 kHz においてフレーム長が 320 点 (20 ms), アルゴリズム遅延が 40 ms であり, AMR-WB+の遅延 72 ms よりも低く設計した. 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, モノラル化し, 16 kHz にダウンサンプリングした後 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包絡の解像度伸縮有り・無しの TCX, 及び参照方式として AMR-WB+の計 3 方式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, 解像度伸縮包絡を導入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 曲全てにおいて解像度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. この結果は, 多くの自然音においてエネルギーの偏りが低域に見られることに起因し, エネルギーの偏りに合わせて包絡の解像度にも偏りを持たせることで情報抽出の効率が上げられることを示唆する.

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非負値疎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, メル対数的に伸縮した場合は 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, その伸縮の度合いに応じて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, 解像度伸縮が実際に符号化器の品質を向上させ得ることを統計的に示した. 今回の実験ではメル対数周波数軸への伸縮のみを使用したが, 符号化対象の信号に合わせた適切な伸縮を探索する方法を考えることが今後の課題である.

謝辞

本研究は JSPS 科研費 26730100 の助成を受けたものです。

参考文献

[1] 3GPP TS 26.290 version 11.0.0 Release 11, 3GPP, 2012.
[2] M. Neuendorf, et al., ”MPEG Unified Speech and Audio Coding - The ISO/MPEG Standard for High-Efficiency Audio Coding of All Content Types”, AES 132nd Convention, Budapest, HU, Apr. 2012.
[3] G. Fuchs, et al., ”MDCT-Based Coder for Highly Adaptive Speech and Audio Coding,” EUSIPCO., IEEE, pp.1264-1268, 2009.
[4] R. Rice and J. Plaunt, ”Adaptive Variable-Length Coding for Efficient Compression of Spacecraft Television Data,” Transaction on Communication Technology, IEEE, Vol. COM-19, No.6, Dec., 1971.
[5] 今井 聖, et al, ”音声合成のためのメル対数スペクトル近似 (MLSA) フィルタ,” 電子通信学会論文誌 ’83/2, Vol.J66-A, No.2, pp.122-129, 1983.
[6] K. Koshida, et al., ”Efficient Encoding of Mel-Generalized Cepstrum for CELP Coders,” ICASSP-97., IEEE, Vol.2, pp.1355-1358, 1997.
[7] K. Koshida, et al., ”A Wideband CELP Speech Coder at 16 kbit/s Based on Mel-Generalized Cepstral Analysis,” ICASP-98, IEEE, Vol. 1, pp.161-164, 1998.
[8] 板倉文忠, ”統計的手法による音声分析合成に関する研究,” 博士論文, 名古屋大学大学院工学研究科, 1972.
[9] 亀岡 弘和, et al., “スペクトル制御エンベロープによる混合音中の周期および非周期成分の選択的イコライザ,” 情報処理学会研究報告, vol. 2006-MUS-66, pp.77-84, Aug. 2006.
[10] http://sp-tk.sourceforge.net/ (’13 年 1 月現在).