2010 年度第 6 回情報処理学会東北支部研究会 (山形大学)　資料番号　10-6-B2-1

仮想音響空間内の音声了解度推定に用いるひずみ尺度の検討*

○小林洋介, 近藤和弘 (山形大)

1　はじめに

人間の両耳効果に基づく聴覚ディスプレイ [1] を用いた音響システムが登場してきた．これまで聴覚ディスプレイは高臨場感再生法として用いられてきたが，拡張音響現実 (Augmented Audio Reality:AA R) システムにおける情報オブジェクトにも利用可能である．AAR システムの評価を考えた時，付加した音声情報の臨場感も重要であるが，システム使用者の主観音声品質の評価が必要である．特に音声システムにおいては，聞き取りやすさの主観評価法である明瞭度・了解度が重要な指標となる．

我々はこれまで AAR システムを志向した頭部伝達関数 (Head-Related Transfer Functions:HRTF) を用いた仮想音響空間内の立体音声の音声了解度を評価してきた [2] その結果，聴取者の正面に定位した話者音声から妨害雑音を水平面で 45 deg. 以上話して定位すると音声了解度が向上することが確認された．またこの時の HRTF による音声了解度の向上分は SNR(Signal to Noise Ratio) で約 6 dB であった [3].

しかし，このような音声了解度の主観評価は評価音が膨大になり，被験者一人あたりの負担が大きくなる．そこで何らかの客観評価指標を用いて音声了解度を推定する手法が必要になる．我々はこれまで ITU-T 勧告 P.862 の PESQ[4] を用いた音声了解度の推定 [5] を行った. その結果 SNR が 10 dB 以上では相関が高かったが，音声了解度の変化が大きい 0 dB から-15 dB にかけては相関は低い結果になった．しかし，別の了解度試験コーパスである HINT[6] を用いた立体音声の主観結果推定では相関係数 r=0.91 と高い値が報告されている [7]．これらの結果から，仮想音響空間内の音声了解度の評価に PESQ を含む何らかのひずみ尺度を用いることが有効であると考えられる．

よって本稿では，我々がこれまで検討してきた仮想空間内音声了解度試験の結果 [3, 8, 9] を用いて仮想音響空間内の音声了解度推定を行い，最適なひずみ尺度を検討する．

2　音声了解度試験

2.1　日本語版 Diagnostic Rhyme Test (DRT)

DRT とは語頭 1 音素のみ異なる単語対の評定用リストを用いて行う了解度試験法である [10]. 被験者は単語対の内の 1 単語のみを聴取し, どちらの音声が聴こえたか二者択一で選択する．評定に用いる単語対の語頭子音は下記の 6 つの属性から成り，これらの単語対を評定することで属性ごとの了解度を測定することができる．以下に 6 つの属性の特徴を示す．属性名の前のラベルは結果の表に対応している．評価単語数は各属性共に 20 単語で総数は 120 単語になる．

(V) Voicing: 有声音と無声音の分類
(N) Nasality: 鼻音と口音の分類
(Su) Sustention: 継続性のある音とない音の分類
(Si) Sibilation: 波形の不規則性に関する分類
(G) Graveness: 抑音と鋭音の分類
(Co) Compactness: スペクトル上のエネルギーが一つのフォルマント周波数に集中するか、分散するかの分類
(A) 120 単語平均: 上記 6 属性の平均値

被験者の正答率を 6 種の音素特徴別，あるいは全回答数の平均で評価する．正答率は式 (1) の調整式により偶発的正答を排除する．

$$S = \frac{R - W}{T} \times 100[\%] \tag{1}$$

ここで S：調整後正答率，R：正答数，W：誤答数，T：全試行数である．これは被験者が全くでたらめに回答した場合に $R \approx W$ となり， $S \approx 0$ となる．

2.2　音像の配置

本稿における音像配置図を Fig.1 に示す．音像の配置には KEMAR-HRIR[11]. を用いた．まず聴取者を中心に，正面を 0 deg., 背面を 180 deg. として右回りを+，左回りを-とする 2 次元の極座標をとり，正面に 0 deg. には評価音を発生する話者音像を定位し，そこから 45 deg. ごとに 8 方位にノイズを定位する．この時の円の半径は KEMAR-HRIR を計測した 1.4 m となる．評価音は KEMAR-HRIR のサンプリング周波数である 44.1 kHz になるように DRT 評価音声と妨害雑音をアップサンプリングして使用した．

2.3　妨害雑音の設定と主観評価

評価音は女性 1 話者の 120 単語にである．妨害雑音は [3, 8, 9] の検討で用いたバブルノイズ [12] を用いた．また，妨害雑音の音圧は，0 deg. において評価音声との SNR が 6，0，-6，-12 dB となるように設

*Distortion measures used for estimation of Japanese speech intelligibility in virtual acoustic space by KOBAYASHI, Yosuke, KONDO, Kazuhiro (Yamagata Univ.)

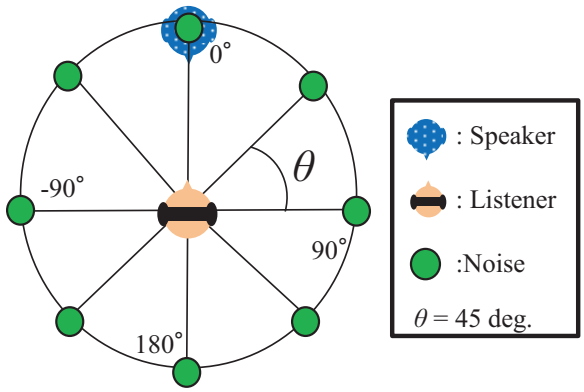

Fig. 1　音像配置図

定した．妨害音の定位方位は Fig.1 に示した 8 方位であり，この各点から 4 種の SNR で再生した．評価音声はすべてヘッドホンで提示した．
被験者数は [3, 8, 9] の結果をまとめたものであり，のべ 28 人のバブルノイズによる主観評価の平均値になる．

3　ひずみ尺度

3.1　セグメンタル SNR [13]

セグメンタル SNR(以下 SNRseg) は時間領域における波形のひずみの大きさを表す尺度の一つで，分析フレームごとの SNR を算出し全フレームの SNR の平均値を用いるひずみ尺度であり，以下の式 (2) で定義される．ここで $x(n)$,$\hat{x}(s)(n)$ は j 番目の分析フレームでの音声信号と雑音重畳音声信号 (劣化音) であり，L は分析フレーム長，M は全フレーム数を示す．

SNRseg=

$$\frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1}x(n)^2}{\sum_{n=Nm}^{Nm+N-1}\left(x(n)-\tilde{x}(n)\right)^2} \quad (2)$$

3.2　周波数重み付セグメンタル SNR[13]

周波数重み付セグメンタル SNR(以下 fwSNRseg) は評価信号をフレームで切り出したのち，さらに帯域ごとに分割して各帯域ごとの重み係数をかけて平均をとったものである．重み係数は人間の主観値への対応が良くなるように [14] で定められている．fwSNRseg の算出式を式 (3) に示す．ここで，$W(j,m)$ は帯域 j の重み係数，$X(j,m)$ と $\hat{X}(n)$ は無劣化音声と雑音重畳音声の m 番目のフレームの帯域 j，K は分析帯域番号，M は全フレーム数を示す．

fwSNRseg=

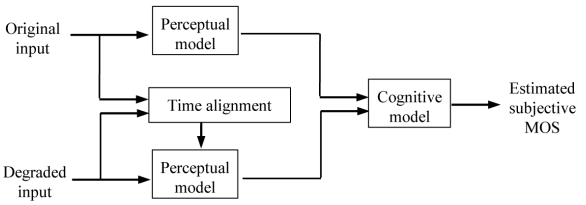

Fig. 2　PESQ アルゴリズム

$$\frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K}W(j,m)\log_{10}\frac{X(j,m)^2}{\left(X(j,m)-\hat{X}(j,m)\right)^2}}{\sum_{j=1}^{k}X(j,m)^2} \quad (3)$$

3.3　LAR 距離 [13]

対数断面積比距離 (Log-Area Ratio distance 以下 LAR) は，LPC 係数をもちいた評価量であり，声道を音響管モデルとしてとらえた時の反射係数を用いる．LAR を求めるための LAR パラメータ $g(j)$ は式 (4) になる．ここで P は LPC 次数であり，η_j は反射係数である．

$$g(j)=\frac{1}{2}\log\frac{1+\eta_j}{1-\eta_j}=\tanh^{-1}\eta_j \quad (4)$$

この LAR パラメータを原信号と劣化信号の両方で求めそれらの差の全フレームでの平均は式 (4) になる．ここで $g_x(j,m)$ と $\hat{g}_x(j,m)$ は原信号と劣化信号の LAR パラメータである．LAR は無劣化音声と雑音重畳音声との差なので値が小さいほど原信号に近い値となる．

$$\mathrm{LAR}=\frac{1}{M}\sum_{m=1}^{M}X(j,m)^2\sqrt{\frac{1}{P}\sum_{j=1}^{P}\lfloor g_x(j)-\hat{g}_x(j)\rfloor} \quad (5)$$

3.4　PESQ[4]

PESQ(Perceptual Evaluation of Speech Quality) は ITU-T 勧告 P.862 で定義されている音声の客観品質評価方式の一つで，ITU-T P.800 勧告で定義される 5 点満点の品質評価法である MOS(Mean Opinion Score) 主観評価 [15] との対応が良い客観評価法である．PESQ 値の算出過程を Fig.2 に示す．無劣化音声と雑音重畳音声 (劣化音声) を知覚モデルを用いて時間，バークスペクトル領域のセルにマッピングする．次にセル間のひずみをバークスペクトルひずみのラウドネスとして算出し，認知モデルを用いて主観 MOS の推定値 (PESQ 値) を得る．

3.5　SNRloss[16]

SNRloss は Jianfen Ma, Philipos C. Loizou が提案している音声了解度の客観評価指標である．SNRloss

は通常の fwSNRseg と，音声強調処理を施した信号との比を 0〜1 の値で表現したもの．0 が劣化がない状態で 1 に近づくほど劣化が大きい．分析時の重み係数には fwSNRseg と同じ重み係数 [14] を用いる．

4 主観評価値とひずみ値

4.1 Better ear モデル

前章のひずみ尺度を用いて，被験者の正面に評価音声を定位しただけの無劣化音声とそれに雑音を重畳した主観音声了解度評価信号のひずみ値 (各尺度のスコア) を求める．しかし，各ひずみ尺度は電話用コーデックの評価用に開発されたものであり，本研究で用いた立体音声（バイノーラル音声）にそのまま用いることはできない．この様な場合，両耳のスコアの平均をとる場合 (Mean ear) と，スコアの良い方つまりひずみの少ない方の耳の値をとる場合 (Better ear) の二つが考えられる．本稿では [7] と同様に Better ear のスコアを用いる．

4.2 ひずみ値の相関

音声了解度の主観評価値とひずみ値 6 の相関を以下の式によって求める．r に用いた $x(n)$ と $d(n)$ は主観評価値とひずみ値の特定のサンプルであり，$\overline{x}$と$\overline{d}$は全サンプルの相加平均を示す．

$$r_s = \frac{\sum (x(n)-\overline{x})(d(n)-\overline{d})}{\sqrt{\sum (x(n)-\overline{x})^2}\sqrt{\sum (d(n)-\overline{d})^2}} \qquad (6)$$

相関係数を Table 1 に示す．表中のラベルは 2.1 に従う．ここで LAR と SNRloss がすべて負の値になっているのはひずみ尺度の定義によるもので，ここでは絶対値を比較する．属性ごとに比較すると，どの尺度も Voicing はやや相関が低く，Sibilation はほぼ無相関である．これは Sibilation は白色雑音に近い特徴の子音であり，バブルノイズなどの環境雑音にもともと頑強であることが原因である [10]．Nasality, Graveness，Compactness は尺度ごとの差が小さく相関が高い．Sustention に関してはひずみ尺度の差がみられ，SNRseg と PESQ は相関が高いものの，fwSNRseg,LAR,SNRloss は極端に低い．120 単語の平均値は全体的に相関が 0.8 以上と高い．

5 音声了解度推定

5.1 了解度推定法

了解度推定のために縦軸に了解度，横軸にひずみ値をとった散布図を作成し，最小二乗法による曲線あてはめを行い，DRT の 6 属性と 120 単語の平均値の合計 7 本の了解度推定関数を求める．この了解度推定関数に，ひずみ値を代入することで了解度を推定する．

Table 1　主観評価値とひずみ値の相関係数

	SNRseg	fwSNRseg	LAR	PESQ	SNRloss
V	0.43	0.41	-0.33	0.38	-0.35
N	0.75	0.74	-0.70	0.72	-0.70
Su	0.91	0.10	-0.05	0.85	-0.06
Si	0.10	0.09	-0.08	0.07	-0.08
Gr	0.88	0.88	-0.82	0.82	-0.83
Co	0.91	0.90	-0.83	0.86	-0.84
A	0.88	0.86	-0.81	0.82	-0.83

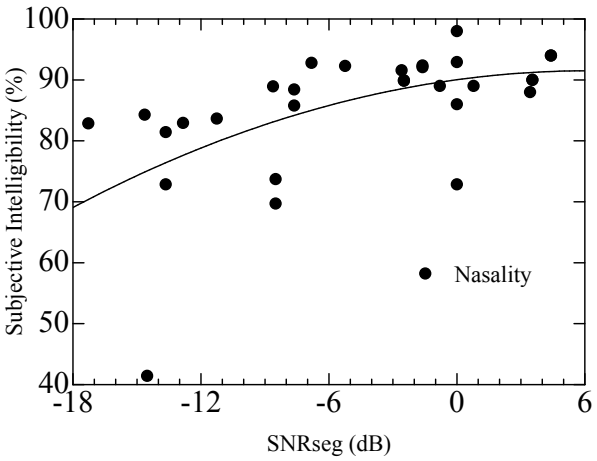

Fig. 3　SNRseg による Nasality の推定曲線

SNRseg を用いたときの Nasality のひずみ値と主観評価値の分布と推定関数を Fig.3 に示す．

5.2 評価指標

推定した了解度の精度推定のために主観評価による了解度との平均二乗誤差 (RMSE) と相関係数 (r) を以下の式 (7) と式 (8) で算出した．r に用いた $x(n)$ と $y(n)$ は主観評価値と推定値の特定のサンプルであり，$\overline{x}$と$\overline{y}$は全サンプルの相加平均を示す．

RMSE=

$$\sqrt{\frac{\sum (\text{Subjective score} - \text{Estimated score})^2}{N}} \qquad (7)$$

$$r_e = \frac{\sum (x(n)-\overline{x})(y(n)-\overline{y})}{\sqrt{\sum (x(n)-\overline{x})^2}\sqrt{\sum (y(n)-\overline{y})^2}} \qquad (8)$$

5.3 推定性能評価試験

音声了解度を推定した結果を Table 2 と Table 3 に示す．表中のラベルは 2.1 で示した内容に一致する．まず音素特徴別にみた場合は，Sibilation とそれ以外の 5 種に分けることができる．Sibilation に関しては

Table 2　性能推定結果 (RMSE)

	SNRseg	fwSNRseg	LAR	PESQ	SNRloss
V	3.18	3.18	4.23	3.64	3.53
N	6.19	5.82	7.32	7.17	6.65
Su	4.83	1.92	2.04	6.55	1.94
Si	2.01	1.93	2.02	2.03	1.93
Gr	5.25	5.17	8.24	8.86	5.82
Co	4.17	3.49	8.43	6.44	4.59
A	3.15	2.31	4.90	4.73	2.90

Table 3　性能推定結果 (相関係数)

	SNRseg	fwSNRseg	LAR	PESQ	SNRloss
V	0.80	0.80	0.60	0.72	0.74
N	0.89	0.90	0.84	0.85	0.87
Su	0.95	0.35	0.05	0.91	0.32
Si	0.24	0.34	0.15	0.12	0.33
Gr	0.96	0.96	0.91	0.88	0.95
Co	0.97	0.98	0.89	0.94	0.97
A	0.97	0.98	0.90	0.91	0.97

RMSE は小さいものの，相関も極端に悪い．これは前章のひずみ値との相関の傾向と同様であると考えられる．その他の音素特徴に関してはひずみ尺度ごとの傾向差はあるが属性による特徴は見られない．
次にひずみ尺度別に見た場合，fwSNRseg の RMSE はすべての属性で一番小さい，しかし Sustention の時の相関は他の尺度と比べて低い値 (r=0.35) になる．SNRseg は fwSNRseg に次いで RMSE が小さいことが多く相関も Sibilation 以外は r=0.80 以上と全体的に高い．LAR は Graveness と Compactness の RMSE が 8% 以上と大きくなり Voicing の相関係数が r=0.60 と若干低く，Sustention についてはほぼ無相関である．PESQ は他と比べると Sustention，Graveness，Compactness の RMSE が大きいが相関はおおむね高い．SNRloss も fwSNRseg と同等の RMSE で相関も高いが，fwSNRseg と同様に Sustention の相関が目立って低い (r=0.32) ．全 120 単語での推定結果を Fig.4 に示す．

6　まとめ

AAR システムを想定した仮想立体音声の音声了解度主観評価の推定に用いるひずみ尺度を検討した．主観評価値とひずみ尺度の相関を比較した結果，どの属性も Voicing はやや相関が低く，Sibilation は無相関であった．Sustention はひずみ尺度ごとの傾向差が大きく，SNRseg と PESQ は相関が高かったものの，他の尺度は無相関に近い値となった．その他の属性はおおむね相関は高く，全単語平均値も高かった．

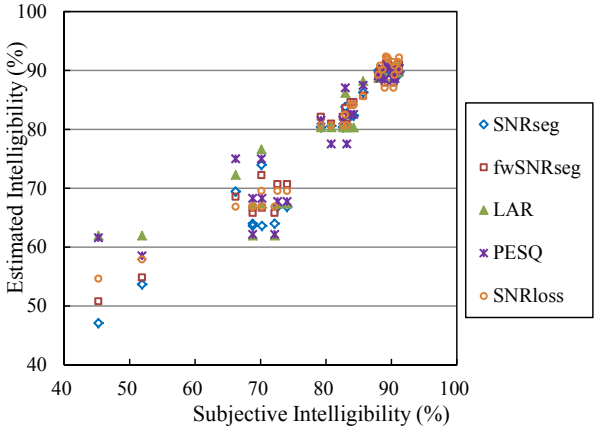

Fig. 4　性能推定結果 (120 単語平均)

また主観値の推定精度も主観値とひずみ値の相関と同傾向だが，Voicing の相関が高くなっているため，主観値とひずみ値の相関の低さが必ずしも推定値の推定精度に影響を与えるとは言い難い．
今後はより推定性能を上げるため，複数のひずみ尺度を併用した評価指標 (一例として [17] など) や fwSNRseg の重み係数を日本語 DRT に対応するように若干変更することなどが考えられる．

参考文献

[1] 鈴木他，信学誌，93(5)，392-396，2007.
[2] Y. Kitashima *et al.*，AST,29(1)，74-81, 2008.
[3] 小林他，音講論 (秋)，735-738, 2011.
[4] ITU-T Recommendation P.862,2001
[5] R. Kaga *et al.*，*Proc. 4th Joint meeting ASA and ASJ*, 3255，2006.
[6] M.J.Nilsson *et al.*，*J. Acoust. Soc. Am.*, 1085-1099，1994.
[7] J. G. Beerends *et al.*，*Proc. Workshop MESAQIN*, On-line，2004.
[8] 矢野他，情処東北研，B-2-3，2008.
[9] 神田他，電学東北連大，2F06，2010.
[10] 近藤他，音響誌，63(4)，196-205，2007.
[11] http://sound.media.mit.edu/resources/KEMAR.html
[12] Rice Univ. Signal Processing Information Base (SPIB) http://spib.rice.edu/
[13] J.R.Deller *et al.*，“Discrete-Time Processing of Speech Signals”, Macmillan，1993.
[14] ANSI Technical Report S3.5, 1997
[15] ITU-T Recommendation P.800,1996
[16] J. Ma *et al.*，Speech Comm.，doi:10.1006/j.specom.2010，2010.
[17] J. Ma *et al.*，*IEEE Trans. ASLP*, Vol.16(1)，2008