E-020

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
On distortion measures for estimation of localized Japanese speech intelligibility in virtual acoustic space

小林洋介 † Yosuke Kobayashi　　**近藤和弘** † Kazuhiro Kondo

1. まえがき

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，拡張音響現実システムの情報付加にも利用可能である．このようなシステムの評価を考えた時，付加情報の臨場感も重要であるが，聞き取りやすさの主観品質である明瞭度・了解度が重要な指標となる．

拡張音響現実の例の一つに小型の端末とイヤフォンを併用したナビゲーションシステムが考えられる．ナビシステムを考えた時，案内に用いる音声そのものの劣化よりも，システムの使用環境によるナビ音声の妨害の影響が大きいと考えらる．特に従来の音声通信では考えられなかった複雑な音場や騒音環境での使用が考えられる．我々は頭部伝達関数 (Head-Related Transfer Functions:HRTF) を用いたバイノーラル音声の了解度を評価し，話者から妨害雑音を離して定位すると了解度が向上することを確認した [2]．しかし，このような了解度試験は評価音数が膨大になり，被験者一人あたりの負担が大きく，客観指標を用いた了解度推定が必要になる．

ダイオティック系では，ひずみ尺度による評価値と事前に分かっている特定の条件の了解度から他の条件の推定了解度を得る方法がある．我々はこれまでダイオティック系で，ITU-T 勧告 P.862 の PESQ[3] を用いた推定では SNR が 0 dB から-15 dB にかけて了解度との相関は低いこと [4] を確認した．他方，Liu らは PESQ やセグメンタル SNR など複数のひずみ尺度による推定了解度結果を比較し，重み付スペクトル包絡距離 (d_{WSS}) が複数の条件で最も相関が高いこと [5]，Beerends らは立体音声の主観評価環境と推定モデルが一致する場合に PESQ を用いた推定では，後述する better ear モデルを使い相関が 0.91 と高いことを報告している [6]．

これらの結果から，バイノーラル系の音声了解度推定にもひずみ尺度を用いることが有効であると考えられる．しかし，バイノーラル音声を評価するのにあたり従来のひずみ尺度には，空間定位による信号歪みをパラメータとして持たないことが問題となる．本論文では，バイノーラル音声の了解度推定に better ear モデルを使い，[2] と同じシステムにおける妨害雑音が異なる場合の了解度推定に最適なひずみ尺度を検討する．

†山形大学大学院理工学研究科, Graduate School of Science and Engineering, Yamagata University

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を聴取し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表 1 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了解度を測定することができる．評価単語数は各属性共に 10 単語対 20 単語で，本稿では 6 属性の平均値を了解度として用いる．正答率は式 (1) の調整式により偶発的正答を排除する．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．

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

表 1: DRT 評価単語の属性

属性名	説明	単語対例
Voicing	有声音と無声音	サイ-ザイ
Nasality	鼻音と口音	マン-バン
Sustention	継続性のある音とない音	ハシ-カシ
Sibilation	波形の規則性	ジャム-ガム
Graveness	抑音と鋭音	ワク-ラク
Compactness	一つのフォルマントへエネルギーが集中するか	ヤク-ワク

2.2. 主観評価モデル

図 1 に本論文で用いた主観評価モデルの音像を示す．各音像は KEMAR ダミーヘッドの HRTF[11] を用いて，聴取者を中心とする水平面上の正面を 0 deg. とし，そこに DRT 評価音を発声する話者音像を定位した．また，0 deg. から 45 deg. ごとの 8 方位にノイズを定位する．これは正面から流れる音声ナビゲーションをさまざまな方位から妨害することを模擬している．また，後述する妨害雑音も雑音抑圧などをして主音声を聞き取りやすくすることではなく，主音声の他にある程度聴取できることを目指している．

表 2 に主観評価実験の諸設定と被験者数の組み合わせを示す．表中のラベルは各主観評価環境を識別するもので，“話者種類-雑音種” となっている．JDRT 評価音声は女性と男性それぞれ 1 話者で，F1 と M1 で区別する．妨害雑音にはバブルノイズ [12]，白色雑音と電子協騒音データベース [13] から幹線道路，在来線電車走行音の環境雑音 2 種の計 4 種を用いた．環境雑音のうち Highway と Railway は評価単語の再生時に自動車と電車が通り過ぎるところになるように切り出して使用した．白色雑音はこれらの環境雑音と極端にスペクトルが異なる雑音を想定している．妨害雑音の音圧は，0 deg. に定位したときに評価音声との SNR を表中の組み合わせで設定した．

F1-B は [7, 8, 9] の主観評価結果のうち，本稿の実験系と重なる結果を抜き出してまとめたもので，のべ 28 人分の主観値になる．各主観評価とも評価単語数は 120 単語× SNR(4 種) ×雑音方位 (8 方位) の 3840 単語になり，雑音重畳した

評価音を被験者にヘッドホンで提示する．各主観評価の被験者は両耳の聴力が健常な 20 代の男女で一部重複する．

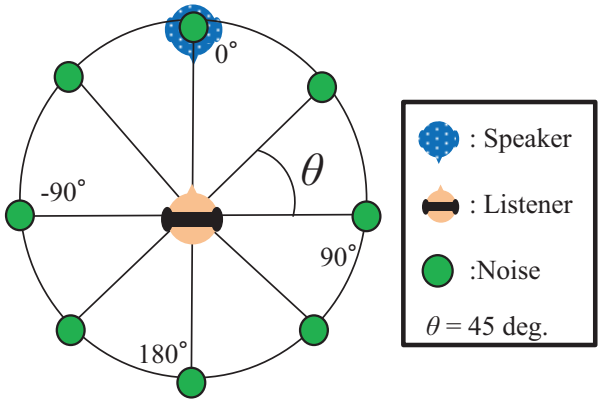

図 1: 音像配置図

表 2: 主観評価の設定

Label	話者	雑音種	SNR (dB)	被験者数
F1-B	F1	Babble	6，0，	のべ 28
F1-W		White	-6，-12	4
F1-H		Highway		6
F1-R		Railway		5
M1-B	M1	Babble		7
M1-W		White		5
M1-H		Highway		6
M1-R		Railway		6

3. 音声了解度推定

3.1. ひずみ尺度を用いた了解度推定法

前章の主観評価の結果をひずみ尺度を用いて推定する．図 2 に了解度推定の流れを示す．まず，評価音声と妨害雑音を評価する音圧に揃え，頭部伝達関数により音像定位した主観評価音を作成する．次にこれらの主観評価音と，JDRT 評価単語の原音を参照信号としてひずみ値を求める．本稿では定位による影響や重畳される妨害雑音を全てを音声のひずみとし，空間定位された主観評価音と了解度試験単語の原音を比較した．

バイノーラル音は人間の両耳受聴を利用した立体音声であるため，左右の両耳に入力される音が異なる．そこで J.G. Beerends らと同様に [6] 左右で劣化の少ない方を用いる．最後にひずみ値から了解度推定関数により客観了解度を得る．了解度推定関数は，主観評価モデル F1-B の了解度とその音質評価値を用いて，シグモイド曲線のあてはめで事前に求めておく．推定関数の例として SNRseg を用いたときのひずみ値と主観値の分布を図 3 に示す．

3.2. ひずみ尺度

3.2.1. セグメンタル SNR [14]

セグメンタル SNR(以下 SNRseg)[14] は時間波形のひずみを表す尺度の一つである．分析フレームごとの SNR を全フレームで平均する指標で，以下の式 (2) で定義される．ここで $x(n)$，$\hat{x}(n)$ は n 番目の分析フレームでの音声と雑音重畳音声 (劣化音) であり，N は分析フレーム長で 20[msec] に設

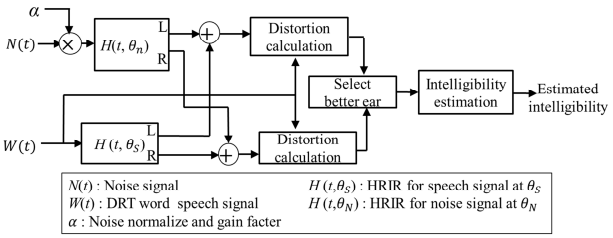

図 2: 了解度推定の流れ

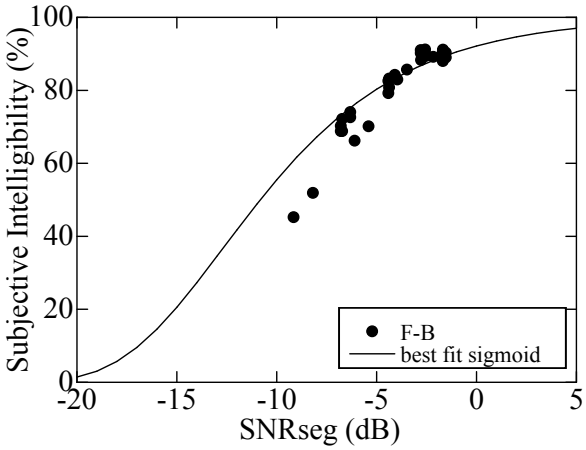

図 3: SNRseg による推定関数

定した，M は全フレーム数を示す．

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

3.2.2. 周波数重み付セグメンタル SNR[14]

周波数重み付セグメンタル SNR(以下 fwSNRseg)[14] 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に重み係数をかけて平均を取ったものである．重み係数は人間の主観音質と対応が良くなるように定められている．fwSNRseg の算出式は式 (3) で定義される．ここで，$X(j,m)$ と $\hat{X}(j,m)$ 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，$W(j,m)$ は帯域 j の m 番目のフレームの重み係数，K は分析帯域数，M は全フレーム数を示す．分析フレーム長は SNRseg と同じ 20[msec] とし，各フレームごとの値は-15〜15[dB] に制限した．

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

3.2.3. ケプストラム距離 [14]

ケプストラム距離 (以下 d_{Cep})[14] は周波数領域のスペクトルひずみの尺度であり，式 (4) で定義される．ここで $c(k)$ と $\hat{c}(k)$ は原信号と劣化信号のケプストラム係数，P は LPC 係数の次数で，本稿では 16 次を用いた. 分析フレーム長は 20[msec] とし，d_{Cep} のスコアの範囲は 0〜10 に制限した．

$$d_{Cep} = \frac{10}{\log 10}\sqrt{2\sum_{k=1}^{P}\left[c(k) - \hat{c}(k)\right]} \tag{4}$$

3.2.4. 対数断面積比距離 [14]

対数断面積比距離 (Log-Area Ratio distance　以下 d_{LAR}) は，LPC 係数を用いた評価量であり，声道を音響管モデルとした時の反射係数を用いる．d_{LAR} を求めるための LAR パラメータ $g(j)$ は式 (5) で定義する．ここで η_j は反射係数である．原信号と劣化信号の両方の LAR パラメータを求め，それらの差を全フレームで平均すると式 (6) 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，M はフレームの総数，P は LPC 係数の次数で本稿では 1 次〜16 次を用いた．

$$g(j) = \frac{1}{2}\log\frac{1+\eta_j}{1-\eta_j} = \tanh^{-1}\eta_j \tag{5}$$

$$d_{LAR} = \frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{P}\sum_{j=1}^{P}\lfloor g_x(j) - \hat{g}_x(j)\rfloor} \tag{6}$$

3.2.5. 周波数重み付スペクトル傾斜距離 [14]

周波数重み付スペクトル傾斜距離 (Weighted Spectral Slope distance 以下 d_{WSS})[14] は d_{Cep} と同様に周波数領域の劣化の評価量であり，式 (7) で定義される．ここで $S(j,m)$ と $\hat{S}(j,m)$ は m 番目のセグメントの j 番目の帯域での原信号と劣化信号の LPC スペクトル傾斜であり，$W(j,m)$ は m 番目のセグメントの j 番目の帯域での重み係数で [15] の値を用いた．M はセグメントの総数, K は分析する帯域の総数で，本稿では K=25 とした．

$$d_{WSS} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K}W(j,m)\left\{S(j,m) - \hat{S}(j,m)\right\}^2}{\sum_{j=1}^{K}W(j,m)} \tag{7}$$

3.2.6.PESQ[3]

PESQ(Perceptual Evaluation of Speech Quality) は ITU-T P.862 で定義される音声品質評価法の一つで，ITU-T P.800 で定義される主観品質評価である MOS(Mean Opinion Score) 評価 [17] と対応する．PESQ 値の算出過程を図 4 に示す．無劣化音声と雑音重畳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し，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 4.5〜-0.5 の範囲で得る．

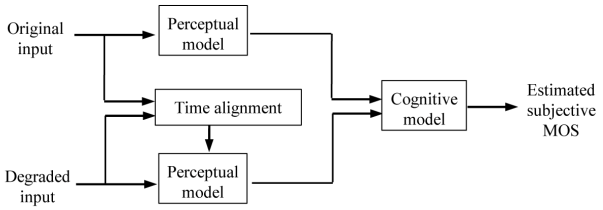

図 4: PESQ アルゴリズム

4. 推定実験

4.1. 評価指標

全 8 モデルの主観評価値のうち，F1-B を用いて求めたシグモイドの推定関数を使い推定値を求めたいモデルの主観了解度推定値を比較する．推定値の精度には，主観値と推定値との平均二乗誤差 (RMSE) とと相関係数 (r) を以下の式 (8) と式 (9) で算出し使用する．ここで，$x(n)$ と $y(n)$ はテストの主観値と推定値のサンプルであり，N は評価総数で評価に用いた SNR と方位角の組み合わせ総数で主観評価条件 1 つにつき 32 になる．

$$\mathrm{RMSE} = \sqrt{\frac{\sum\left(x(n) - y(n)\right)^2}{N}} \tag{8}$$

$$r = \frac{\sum(x(n)-\overline{x})(y(n)-\overline{y})}{\sqrt{\sum(x(n)-\overline{x})^2}\sqrt{\sum(y(n)-\overline{y})^2}} \tag{9}$$

4.2. 推定性能評価 (120 単語平均)

表 3 に全 120 単語平均の尺度ごとの RMSE と相関係数を示す．Label は話者と付加雑音の組み合わせで，表 2 に対応する．Label のうち N_{Op}はノイズオープンテスト, S_{Op}は話者オープンテストに該当するもので, $F1-B$ と雑音種が異なる 6 種と話者が異なる 3 種の条件の $RMSE$ は平均値，相関係数は該当する全ての推定値と主観値の相関係数である．

F1-B 条件を推定した場合 (クローズテスト) は，シグモイドの計算が主観値とひずみ値の間の予測残差を最小にするため，どの条件も RMSE が小さい．この時の RMSE は各ひずみ尺度の推定上限と考えられる．d_{LAR} はこの時ですでに RMSE が 8.25% であり，このほかの推定でも RMSE が大きく相関係数も小さいことから，了解度推定に用いるひずみ尺度としては不適であると考えられる．

他の尺度の傾向は 2 つに分かれる．SNRseg と fwSNRseg の様に主観評価条件のほとんどで RMSE が 10% 以下で，相関係数もほとんどが 0.9 以上になる．もう一方は，d_{Cep},d_{WSS} と PESQ の条件によっては RMSE が小さいときもあるがおおむね RMSE が大きく，相関も下がる物に分けられる．

次に，図 5 と図 6 に SNRseg と PESQ の主観了解度と各尺度の推定結果の一致具合 (推定精度) を示す．各プロット点は同一の SNR で同一のノイズ方位角での各主観評価条件に相当する．黒色のプロットは女性話者の場合で，白抜きは男性話者の場合であり，黒色と白抜きで同一形状のものは同一の雑音種になる．SNRseg は話者や雑音種の影響がほとんど見られず，ほぼすべてのプロットが対角線上にあり精度が高いと言える．PESQ は主観了解度が 60% までは対角線上にあり高い精度であるが，それより低い範囲では対角線から外れ精度が低い．さらに話者が変わった場合は推定了解度が 80% 前後に飽和し，話者依存性がみられる．

5. むすび

拡張音響現実を実現するための立体音声の了解度推定に用いるひずみ尺度を検討した．了解度を推定するための推定関数を導出には我々がこれまでに行ったバブルノイズによる主観評価結果を用いて，主観評価に用いる雑音と話者を変えた場合の値を推定し，主観値と比較した．その結果，SNRseg とそれに重み係数を加えた fwSNRseg を用いた場合は他の尺度よりも推定精度が高かった．また，一部の尺度では推定関

表 3: ひずみ尺度ごとの RMSE(%) と相関係数 r

	RMSE(%)						相関係数 r					
Label	SNRseg	fwSNRseg	d_{Cep}	d_{LAR}	d_{WSS}	PESQ	SNRseg	fwSNRseg	d_{Cep}	d_{LAR}	d_{WSS}	PESQ
F1-B	2.45	1.91	3.17	8.25	4.66	5.42	0.98	0.99	0.97	0.82	0.92	0.89
F1-W	8.35	4.69	26.51	42.60	14.52	14.60	0.93	0.98	0.83	0.54	0.95	0.95
F1-H	6.43	2.95	13.56	12.48	12.44	9.10	0.92	0.97	0.89	0.66	0.91	0.94
F1-R	6.09	7.05	9.30	12.51	12.67	12.64	0.96	0.96	0.87	0.71	0.87	0.78
M1-B	7.93	12.77	18.67	19.12	21.91	20.53	0.98	0.96	0.73	-0.02	0.98	0.93
M1-W	14.27	7.75	23.79	26.62	22.95	22.96	0.97	0.98	0.87	0.63	0.98	0.97
M1-H	3.81	7.08	7.91	13.74	16.50	15.18	0.96	0.98	0.77	0.04	0.98	0.96
M1-R	5.81	12.25	21.29	25.85	26.74	25.91	0.98	0.99	0.76	0.11	0.93	0.92
N_{Op}	7.46	6.96	17.06	22.30	17.64	16.73	0.91	0.80	0.59	0.27	0.43	0.40
S_{Op}	7.95	9.96	17.91	21.33	22.02	21.15	0.95	0.96	0.48	0.22	0.84	0.77

数を導出した主観評価環境と同一の話者でない場合の推定精度が悪くなる．

これらの結果は Liu らの d_{WSS} がおおむね高い相関を示すことなどと傾向が異なる．これは，我々の想定が妨害雑音それ自体を聴こえなくすることではなくいことと，了解度を求めるのに用いたコーパスの特性と考えられる．今後は，複数尺度を混合した音声品質推定モデルや推定関数を導出するモデルの最適化を検討する必要がある．

謝辞

本研究は東北大学電気通信研究所平成 22 年度共同プロジェクト (H21/A10) として行った．関係者各位に感謝する．

参考文献

[1] 鈴木他，“超臨場感音響の展開”, 信学会誌，93(5)，392-396，(2010).
[2] Y. Kitashima *et al.*，“Intelligibility of read Japanese words with competing noise in virtual acoustic space,” Acoust. Sci. Tec. ,29(1)，74-81, (2008).
[3] ITU-T Recommendation P.862, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” Feb. 2001.
[4] R. Kaga *et al.*，“Towards estimation of Japanese intelligibility scores using objective voice quality assessment measures,” *Proc. 4th Joint meeting ASA and ASJ*, 3255，(2006).
[5] W.M.Liu *et al.*，“Assessment of Objective Quality Measures for Speech Intelligibility” *Proc. Interspeech 2008*, 699-702, (2008).
[6] J. G. Beerends *et al.*，“Measurement of speech intelligibility based on the PESQ approach,” *Proc. Workshop MESAQIN*,(2004).
[7] 矢野他，“音像定位した音声と妨害雑音間距離が及ぼす音声了解度への影響”, 情処東北研，2007-6-B-2-3，(2008).
[8] 小林他，“仮想 3 次元空間における音声了解度推定の検討” , 音講論 (秋)，735-738, (2011).
[9] 神田他，“空気伝導と骨伝導ヘッドホンを用いた空間定位音声了解度の比較”, 電学東北連大，2F06，(2010).
[10] 近藤他，“二者択一型日本語音声了解度試験方法の検討”, 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] 日本電子工業振興協会 騒音データベース (JEIDA-NOISE),http://research.nii.ac.jp/src/list/detail.html
[14] J.R.Deller *et al.*，“Discrete-Time Processing of Speech Signals,” Macmillan，(1993).
[15] D.Klatt，“Prediction of perceived phonetic distance from critical band spectra,” *Proc. IEEE ICASSP.*, pp.281-2822, (1982).
[16] ANSI Technical Report S3.5, (1997).
[17] ITU-T Recommendation P.800,“Methods for subjective determination of transmission quality,”(1996).

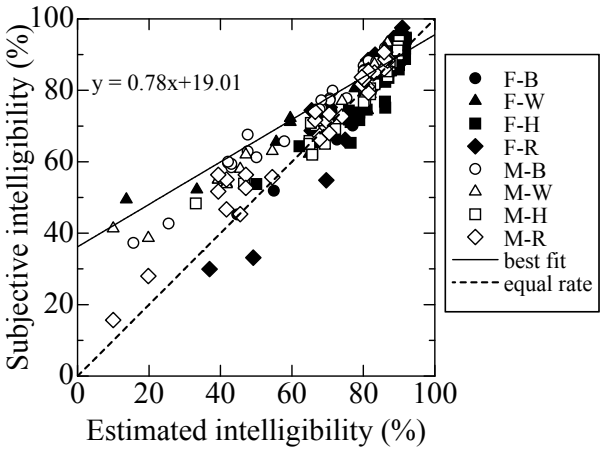

図 5: SNRseg の推定精度

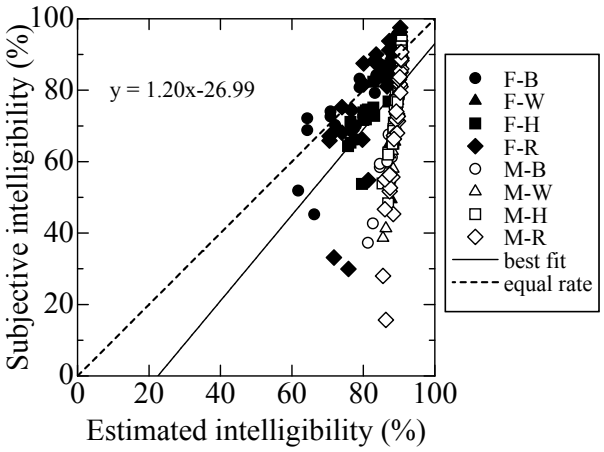

図 6: PESQ の推定精度