音楽情報科学　39－22
（2001．2．23）

雑音下母音聴取における雑音のスペクトル構造の影響

石塚 健太郎　　相川 清明

NTT コミュニケーション科学基礎研究所

〒243-0198 神奈川県 厚木市 森の里若宮 3-1
e-mail: ishizuka@atom.brl.ntt.co.jp, aik@idea.brl.ntt.co.jp

あらまし

高雑音環境下で人間がどのように音韻を知覚するかを調べるため，様々な雑音を単母音に低S/N比で重畳した場合の音韻知覚について2種の聴取実験を行った。まず，雑音として過渡的・定常的な種々の雑音を重畳した刺激音を用い聴取実験を行った。この結果，刺激音のスペクトル包絡が同様の場合にも，より微細な音響特徴により高雑音下での音韻知覚を行っている可能性が示唆された。次に，平坦なスペクトル包絡を持つ雑音の調波構造の違いが音韻知覚に及ぼす影響について調査するための聴取実験を行った。この結果，聴覚が時間周波数領域で必要な音声特徴を取捨選択し利用している可能性，および高雑音下の音声知覚においては音声の偶数倍音成分の寄与が低い可能性が示唆された。

キーワード　雑音環境，音韻知覚，音響特徴，スペクトル構造，聴取実験

Effect of Spectral Structure of Noises on Noisy Vowel Perception

Kentaro Ishizuka　　Kiyoaki Aikawa

NTT Communication Science Laboratories

Morinosato-Wakamiya 3-1, Atsugi City, Kanagawa, 243-0198, Japan
e-mail: ishizuka@atom.brl.ntt.co.jp, aik@idea.brl.ntt.co.jp

Abstract

This paper describes new findings on the effect of the spectral structure of noises on noisy vowel perception. We conducted two experiments to examine how listeners perceive natural vowels under high noisy environmental conditions at around -2 dB in SNR. First, we used periodic/transient environmental recorded noises or artificially generated noises. The result suggests that the human auditory system uses more detailed features of sounds rather than the spectral envelopes to perceive vowels in noisy environments. Second, we used several types of harmonic structured noise that had an identical spectral envelope. The result suggests that the even harmonic component of the vowel make a less contribution to noisy vowel perception than the odd harmonic components. Furthermore, the result suggests that the human auditory system changes dynamically to use time/frequency features corresponding to waveform and spectral structure.

key words　noisy environment, phoneme perception, acoustical feature, spectral structure, hearing experiment

1 はじめに

雑音に頑健な音声認識手法が求められている[1]。本研究では，人間の聴覚機構を分析することで，雑音環境に頑健な音声特徴量を得ることを目的とする。

現在の音声認識手法においては，音声のスペクトル包絡を特徴量とし音声認識を行う手法が主流である[2]。このために，雑音が重畳された音声が入力され，入力音のスペクトル包絡がクリーンな音声のものと大きく異なる場合は認識率が低下する。とりわけ，雑音のスペクトルパワーが音声のスペクトルパワーを大きく上回っている場合は，音声に関わらず入力音のスペクトル包絡が同様のものとなり，音声認識結果は同じものとなってしまう。

一方，人間はそのような場合においても，雑音によっては目的音を分凝し聴き分けることができる[3]。これは，人間がなんらかの手がかりを元に，雑音や目的音に対するスペクトル包絡以外の音響的特徴を獲得しているためと考えられる。

雑音中の音声知覚に関しては，音声明瞭度やマスキング[4][5]，音素修復[6][7]，聴覚フィルタ[8]の研究において多く調べられているが，日常耳にする音や複合音，調波構造を持つ雑音を用いたもに関しては，純音に対するマスキングの研究[9][10][11]はあるものの，音声を対象としたものは少ない。

本稿では，高雑音環境下での人間の聴覚の音韻知覚について検討する。そのために，様々な雑音を単母音に低S/N比で重畳した場合の音韻知覚について2種の聴取実験を行った。まず，雑音として過渡的・定常的な種々の雑音を重畳した刺激音を用い聴取実験を行った。これを以下実験1とする。次に，平坦なスペクトル包絡を持つ雑音の調波構造の違いが音韻知覚に及ぼす影響について検討するための聴取実験を行った。これを以下実験2とする。

以下，2節で実験1について述べ，3節で実験2について述べ，4節でまとめを述べる。

2 実験1

定常的・過渡的な雑音を低S/N比で重畳した単母音を人間が聴いた際，雑音の種類およびS/N比の違いが高雑音下の音韻知覚に与える影響を調査した。

単母音に1種類の雑音を重畳した刺激音を，母音・雑音の種類とS/N比を様々に変化させ作成した。この刺激音を被験者に呈示し，刺激音に含まれる母音の判断を求めた。その上で，雑音の種類ごとに正答率を求めた。

2.1 刺激

母音として，ATR日本語データベース中の女性1名(FAF)により単独発声された/a/, /i/, /u/, /e/, /o/の5母音を用いた。母音の平均パワーについては母音間で同一にした。元データのサンプリング周波数は12 kHzであったが，これを11.025 kHzにダウンサンプリングして使用した。

雑音として，定常的雑音7種類，過渡的雑音6種類を用いた。定常的雑音としては，

(a) 基本周波数100 Hzのブザー音
(b) 基本周波数440 Hzのクラリネット音
(c) 基本周波数440 Hzのパイプオルガン音
(d) 白色雑音
(e) ピンク雑音
(f) 中心周波数・変調幅をランダムに定めFM変調した正弦波を200種類重畳した複合音
(g) 440 Hzの整数倍の周波数を持つ正弦波を13倍音まで同じ強さで重畳した調波音を，-3 dB/oct.の減衰特性を持つフィルタに通した複合音

を用い，過渡的雑音としては，

(h) ガラスの割れる音
(i) 和太鼓の音
(j) シンバルの音
(k) 雑音(h)と同じ振幅包絡を持つ白色雑音
(l) 雑音(i)と同じ振幅包絡を持つ白色雑音
(m) 雑音(j)と同じ振幅包絡を持つ白色雑音

を用いた。雑音(a), (b), (c), (h), (i), (j)については，元データのサンプリング周波数は44.1 kHzであったが，これを11.025 kHzにダウンサンプリングして使用した。その他の雑音はサンプリング周波数11.025 kHzで人工的に作成した。雑音(k)〜(m)における振幅包絡とは，波形領域における振幅の時間変化のことであり，白色雑音の単位時間あたりのパワーを雑音(h)〜(j)と同一にすることにより作成した。

刺激音は，音声のパワーは一定で，上記それぞれの雑音のパワーを変化させS/N比-1.58 dB, -2.28 dB, -2.92 dBで重畳することによって構成した。この結果，刺激音の数は195種類となった。刺激音の持続時間は約0.3秒で，立ち上がりと立下りに10 msのテーパーをかけた。刺激音のサンプリングレートは11.025 kHzであった。

刺激音のセットとして，重畳されている雑音の種類によって以下の3セットを用意し，被験者によって呈示するセットを変えた。

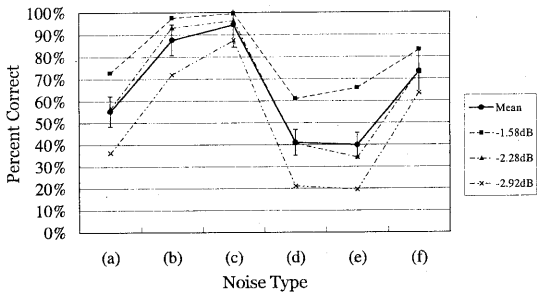

図 1: 刺激音セット 1 の平均正答率と標準偏差

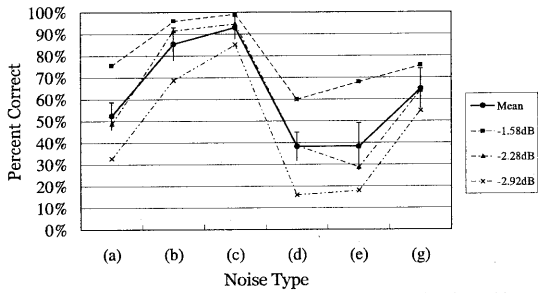

図 2: 刺激音セット 2 の平均正答率と標準偏差

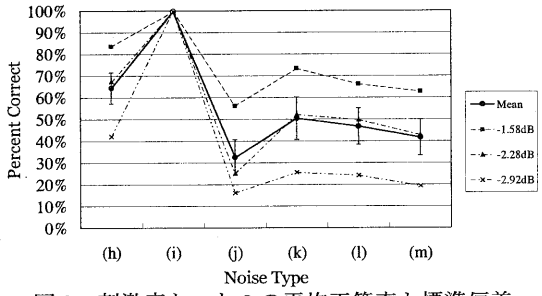

図 3: 刺激音セット 3 の平均正答率と標準偏差

- **刺激音セット 1**: 雑音(a), (b), (c), (d), (e), (f) が重畳された刺激音のセット
- **刺激音セット 2**: 雑音(a), (b), (c), (d), (e), (g) が重畳された刺激音のセット
- **刺激音セット 3**: 雑音(h), (i), (j), (k), (l), (m) が重畳された刺激音のセット

2.2 手続き

刺激音は 6 秒間隔で呈示され，被験者は刺激音に含まれる母音についてマークシートに強制選択で回答した。被験者 1 名に対し，1 種類の刺激音について 5 回呈示したため，刺激音セット 1 つあたり 450 試行となった。実験は，被験者 1 名ごとに，最初に刺激音セット 1 または 2 について行い，15 分の休憩を挟んで刺激音セット 3 について行った。刺激音の呈示順は被験者ごとにランダムに選ばれた。刺激作成と実験制御には計算機を用い，刺激呈示は防音室内でヘッドフォン(STAX SR-□ Signature)を用いて両耳に行った。母音は平均音圧 60 dB SPL で呈示した。

2.3 被験者

18 歳から 54 歳の男女各 10 名，計 20 名。うち男女各 5 名，計 10 名に刺激音セット 1 と 3，他の 10 名に刺激音セット 2 と 3 を用いた。

2.4 実験結果

図1, 2, 3に刺激音セット1, 2, 3についての雑音ごとの平均正答率および標準偏差と，S/N比別の平均正答率を示す。チャンスレベルは20%である。

刺激音セット1

雑音(c)での正答率が最も高く，雑音(e)での正答率が最も低かった。雑音の種類に関わらずS/N比の低下に伴い正答率は低下した。雑音(d), (e)についてはS/N比-2.92 dBにおいてチャンスレベルと同等の正答率しか得られなかったが，その他の雑音については低S/N比であっても36%以上の正答率が得られた。

雑音ごとの正答率間には，分散分析の結果 1% 水準で有意差があった。雑音(b)と雑音(c)，雑音(d)と雑音(e)の間以外の雑音種類間では，Turkey-Kramer の多重比較検定の結果 5%水準で有意な差があった。

刺激音セット2

雑音(c)での正答率が最も高く，雑音(d), (e)での正答率が最も低かった。雑音の種類に関わらずS/N比の低下に伴い正答率は低下した。雑音(d), (e)についてはS/N比-2.92 dBにおいてチャンスレベルと同等の正答率であったが，その他の雑音については低S/N比であっても32%以上の正答率が得られた。

雑音ごとの正答率間には，分散分析の結果 1%水準で有意差があった。雑音(b)と雑音(c)，雑音(d)と雑音(e)の間以外の雑音種類間では，Turkey-Kramer の多重比較検定の結果 5%水準で有意な差があった。

刺激音セット3

雑音(i)での正答率が最も高く，雑音(j)での正答率が最も低かった。雑音(i)以外ではS/N比の低下に伴い正答率が低下した。雑音(i)についてはS/N比が低下しても正答率が変化しなかった。

雑音ごとの正答率間には，分散分析の結果 1%水準で有意差があった。雑音(k)と雑音(l)，雑音(l)と雑音(m)の間以外の雑音種類間では，Turkey-Kramer の多重比較検定の結果 5%水準で有意な差があった。

2.5 考察

刺激音セット 1,2 いずれにおいても，基本周波数が高く音のスペクトルピークが少ない，調波構造を持つ定常的雑音(b), (c)での正答率が有意に高く，スペクトルが平坦な定常的雑音(d), (e)での正答率が有意に低い。このことから，平坦なスペクトルを持つ雑音の音韻知覚妨害効果が高いことがわかる。

一方，調波構造を持つ雑音(a)での正答率は雑音(b), (c)よりも低い。これは，雑音(a)の基本周波数が低く，多くの調波成分を持ち，かつ高周波まで高いスペクトルパワーを持つために高周波領域でのスペクトルパワーが低い雑音(b), (c)よりも周波数領域上での音声特徴知覚を妨害したためと考えられる。

雑音(f)は構成音の数が多く，ほぼ平坦なスペクトルを持つが正答率は高い。これは，構成音が FM 変調されていることによって，共変調マスキング解除[12]と同様の現象によって，雑音のマスキング量が低下したためと考えられる。

刺激音セット 2 の雑音(e)と(g)は同様のスペクトル包絡を持つが，雑音(g)での正答率は有意に高い。このことから，人間がスペクトル包絡よりも細かなスペクトル特徴を用い高雑音下での音韻知覚を行っている可能性が示唆される。

刺激音セット 3 において，雑音(i)は低 S/N 比においても音韻知覚を全く妨害しない。これは，雑音(i)のスペクトルが低周波にのみ集中し，母音の周波数領域での特徴知覚を妨害しないことによると考えられる。

雑音(h)～(j)と雑音(k)～(m)は同様に時間方向のパワーが変化するが，正答率には有意な差がある。このことから，パワーの時間変化が同じ場合でもスペクトル構造が異なれば音韻知覚妨害効果が全く異なることがわかる。その一方で，雑音(k)と(m)についてはスペクトル構造が類似していても正答率に有意な差があり，パワーの時間変化によっても音韻知覚妨害に差が出ることがわかる。

3 実験2

実験 1 の結果より，類似したスペクトル包絡を持つ定常雑音においても，スペクトルの微細構造の違い，また雑音の調波構造の違いによって音韻知覚妨害に差が出ることがわかった。

そこで，本節では平坦なスペクトル包絡を持つ各種の雑音の調波構造が，高雑音下での音韻知覚にもたらす影響について検討する。

雑音の調波構造は母音のフォルマント周波数に関わらず，音声の調波成分に応じて定めた。聴覚末梢・中枢系でのスペクトル処理機構においては，低次の処理系では母音は調波成分が分離されたまま表現され，フォルマント周波数は高次の認知処理部で抽出されていると考えられている[13]。

単母音に 1 種類の雑音を重畳した刺激音を，母音・雑音の種類と S/N 比を様々に変化させ作成した。この刺激音を被験者に呈示し，刺激音に含まれる母音の判断を求めた。その上で，雑音の種類ごとに正答率を求めた。

3.1 刺激

母音としては，実験 1 と同じものを用いた。自己相関法により求めた基本周波数は母音間で 254.3 Hz ～266.3 Hz で平均 259.7 Hz であった。

雑音として，母音の基本周波数に応じ以下の 11 種類を作成し用いた。調波音については，サンプリングレートにより決まる最大周波数まで同じ位相・強さの正弦波を重畳することで構成した。

(A) 白色雑音
(B) 260 Hz の整数倍の正弦波による調波音
(C) 260 Hz の偶数倍の正弦波による調波音
(D) 260 Hz の奇数倍の正弦波による調波音
(E) 130 Hz の整数倍の正弦波による調波音
(F) 130 Hz の整数倍の正弦波による調波音から 260 Hz の奇数倍の成分を除いたもの
(G) 130 Hz の整数倍の正弦波による調波音から 260 Hz の偶数倍の成分を除いたもの
(H) 130 Hz の奇数倍の正弦波による調波音
(I) 220 Hz の整数倍の正弦波による調波音
(J) 240 Hz の整数倍の正弦波による調波音
(K) 280 Hz の整数倍の正弦波による調波音
(L) 300 Hz の整数倍の正弦波による調波音

刺激音は，音声のパワーは一定で，上記それぞれの雑音のパワーを変化させ S/N 比-1.58 dB, -2.28 dB, -2.92 dB で重畳することによって構成した。この結果，刺激音の数は 165 種類となった。刺激音の持続時間は約 0.3 秒で，立ち上がりと立下りに 10 ms のテーパーをかけた。刺激音のサンプリングレートは 11.025 kHz であった。

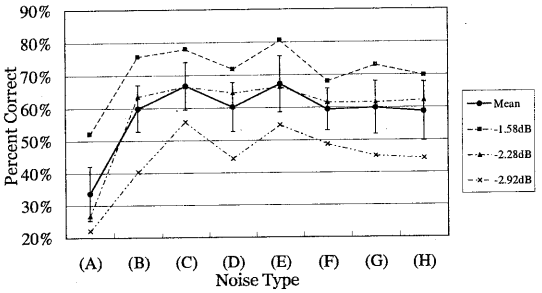

図 4 : 刺激音セット 4 の平均正答率と標準偏差

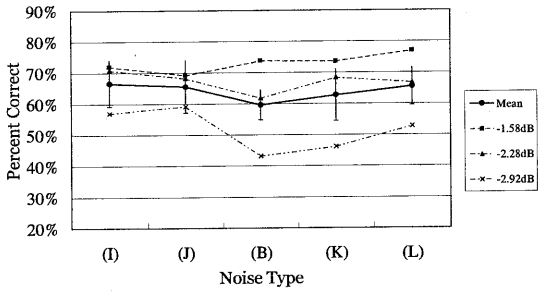

図 5 : 刺激音セット 5 の平均正答率と標準偏差

刺激音のセットとして，重畳されている雑音の種類によって以下の 2 セットを用意した。

- **刺激音セット** 4: 雑音(A), (B), (C), (D), (E), (F), (G), (H) が重畳された刺激音のセット
- **刺激音セット** 5: 雑音(B), (I), (J), (K), (L) が重畳された刺激音のセット

3.2 手続き

刺激音は 6 秒間隔で呈示され，被験者は刺激音に含まれる母音についてマークシートに強制選択で回答した。被験者 1 名に対し，1 種類の刺激音について 5 回呈示したため，刺激音セット 4 で 600 試行，刺激音セット 5 で 375 試行となった。実験は，被験者 1 名ごとに，最初に刺激音セット 4 について行い，15 分の休憩を挟んで刺激音セット 5 について行った。刺激音の呈示順は被験者ごとにランダムに選ばれた。刺激作成と実験制御には計算機を用い，刺激呈示は防音室内でヘッドフォン(STAX SR-□□ Signature)を用いて両耳に行った。母音は平均音圧 60 dB SPL で呈示した。

3.3 被験者

実験 1 とは異なる，19 歳から 65 歳の男女各 10 名，計 20 名。

3.4 実験結果

図 4, 5 に刺激音セット 4, 5 についての雑音ごとの平均正答率および標準偏差と，S/N 比別の平均正答率を示す。チャンスレベルは 20%である。

刺激音セット4

雑音(E)での正答率が最も高く，雑音(A)での正答率が最も低かった。雑音(B)～(H)では雑音(H)での正答率が最も低い。また，雑音の種類に関わらず S/N 比の低下に伴い正答率は低下した。雑音(A)では S/N 比-2.92 dB においてチャンスレベルと同等の正答率であったが，その他の雑音については低 S/N 比であっても 40%以上の正答率が得られた。

雑音ごとの正答率間には，分散分析の結果 1%水準で有意差があった。また，雑音(A)と雑音(B)～(H)，雑音(E)と雑音(F), (H)，雑音(C)と雑音(H)間の正答率については，Turkey-Kramer の多重比較検定の結果 5%水準で有意な差があった。

刺激音セット5

雑音(I)での正答率が最も高く，雑音(B)での正答率が最も低かった。また，雑音の種類に関わらず S/N 比の低下に伴い正答率は低下した。低 S/N 比であってもすべての雑音で 40%以上の正答率が得られた。

雑音ごとの正答率間には，分散分析の結果 5%水準で有意差があった。また，雑音(B)と雑音(I)間の正答率については，Turkey-Kramer の多重比較検定の結果 5%水準で有意な差があった。

3.5 考察

実験の結果，刺激音のスペクトル包絡がほぼ平坦となる場合であっても，雑音のスペクトル構造および調波構造の違いにより正答率に差が出ることがわかった。

刺激音セット 4 において，雑音のスペクトル包絡は全て平坦となるが，雑音(A)での正答率が際立って低く，その一方で雑音(B)～(H)では最悪でも正答率が 40%を下らない。これらの結果より，実験 1 と同様，人間がスペクトル包絡よりも微細なスペクトル構造を知覚できる可能性が示唆される。

雑音(E)での正答率が最も高かった理由としては，130 Hz の整数倍の正弦波を同じ位相・強さで 42 倍音まで重畳したことから雑音を構成する音の数が最も多く，雑音(B)～(H)の中で最もスペクトルパワーのピークが低いために周波数領域での音声特徴が知

覚されやすかった点が考えられる。加えて，雑音(E)は重畳した正弦波の数が多いため，時間領域において 1/130 秒(約 7.6 ms)周期で急激な減衰振動を示す波形であることから，その雑音パルス間において現われる音声波形から音声特徴を知覚したとも考えられる。これらのことから，聴覚にはこのような時間周波数領域で必要な情報を統合・取捨選択する能力があり，その結果正答率が高くなった可能性が示唆される。

また，音声の偶数倍音近辺にのみ雑音の構成音がある雑音(C)での正答率が高く，このことから，高雑音下での音韻知覚においては，音声の偶数倍音成分の寄与が低い可能性が示唆され，基本周波数の同定が重要であると考えることもできる。もしそうならば，雑音(H)を含む刺激音では音声の調波成分が刺激音全体の偶数倍音成分に相当するために音韻知覚が妨げられ，その結果正答率が悪化したと見ることもできる。

また，刺激音セット 5 においては，音声の調波構造に最も近い構造を持つ雑音(B)での正答率が最も低く，雑音の基本周波数が音声の基本周波数より離れるほど正答率が高くなる。このことから，同様の調波構造を持つ雑音であっても，基本周波数が母音の基本周波数に近いほど音韻知覚妨害効果が高い可能性が示唆される。

4 まとめ

本稿では，様々なスペクトル構造と時間変化を持つ雑音を単母音に低 S/N 比で重畳した場合の音韻知覚について聴取実験を行った。その結果，同じスペクトル包絡の雑音でも調波構造の違いにより音韻知覚妨害に差のあることがわかり，高雑音下での音韻知覚について幾つかの示唆を得た。

実験 1, 2 を通じ，雑音のスペクトルが平坦で，調波構造を持たない雑音の音韻知覚妨害効果が高いことがわかる。調波構造を持つ雑音の場合は，構成音数や調波成分のスペクトルパワーが異なる場合に，音韻知覚妨害に差がある場合と無い場合があることがわかった。

また，高雑音下の音韻知覚において，人間がスペクトル包絡よりも微細なスペクトル構造を知覚している可能性についての示唆，および時間周波数領域での音響特徴を取捨選択し統合している可能性についての示唆，音声の基本周波数同定が重要である可能性についての示唆を得た。

以上のように幾つかの可能性を示唆する結果を得たが，本稿の結果だけでは結論が出せない点が多いため，高雑音下における音韻知覚に高い寄与を持つ音響特徴については今後も検討を進める必要がある。

謝辞

日頃研究を支援して頂く NTT コミュニケーション科学基礎研究所メディア情報部萩田紀博部長，活発な議論を頂く対話研究グループの皆様に感謝します

参考文献

[1] 中川聖一, “音声認識研究の動向,” 信学論, Vol.J83-D-II, pp.433-457, 2000.

[2] Hunt,J.M, “Spectral Signal Processing for ASR,” Proc. of ASRU’99 Workshop, 1999.

[3] Bregman,A.S., “Auditory Scene Analysis,” MIT Press, 1990.

[4] Miller,G.A., “The Masking of Speech,” Psychological Bulletin, Vol.44, No.2, pp.105-129, 1947.

[5] Pickett,J.M., “Perception of Vowels Heard in Noises of Various Spectra,” J. Acoust. Soc. Am., Vol.29, No.5, pp.613-620, 1957.

[6] Warren,R.M., Sherman,G.L., “Phonetic restorations based on subsequent context,” Percept. and Psycho., Vol.16, No.1, pp.150-156, 1974.

[7] 柏野牧夫, “閉鎖区間の前後に分散する手がかりに基づく日本語語中閉鎖子音の知覚,” 音響誌, Vol.48, No.2, pp.76-86, 1992.

[8] Ainsworth,W.A., Meyer,G.F., “Recognition of plosive syllables in noise: Comparison of an auditory model with human performance,” J. Acoust. Soc. Am., Vol.96, No.2, pp.687-694, 1994.

[9] Neff,D.L., Green,D.M., “Masking produced by spectral uncertainly with multicomponent maskers,” Percept. Psycho., Vol.41, No.5, pp.409-415, 1987.

[10] Oh,E.L., Lufti,R.A., “Informational masking by everyday sounds,” J. Acoust. Soc. Am., Vol.106, No.6, pp.3521-3528, 1999.

[11] Oh,E.L., Lufti,R.A., “Effect of masker harmonicity on informational masking,” J. Acoust. Soc. Am., Vol.108, No.2, pp.706-709, 2000.

[12] Moore,B.C.J., 大串健吾訳, “聴覚心理学概論,” 誠信書房, pp.120-126, 1994(原著 3rd. Ed. 1989).

[13] 平原達也, “母音知覚と聴覚機構,” 音講論秋季, 2-9-10, pp.365-366, 1991.