非会員 崔 載 承 (大阪市立大学)

正員 岡 本 次 郎 (大阪市立大学)

正員 細 川 省 一 (大阪市立大学)

Improvement of the Characteristic of Speech which is buried in Noise based on Auditory Physiology

Jae-Seung Choi, Non-member, Jiro Okamoto, Member,
Shoichi Hosokawa, Member (Osaka City University)

On the field of speech processing, the treatment of noise is still important problems for speech research. Especially, it has been noticed that the background noise causes remarkable reduction of speech recognition ratio. As the examples of the background noise, there are such various non-stationary noises existing in the real environment as driving noise of automobiles on the road or typing noise of printer. The treatment for these kinds of noises is not so simple as could be eliminated by the former Winer filter, but needs more skillful techniques. In this paper as one of these trials, we show an algorithm which is a speech enhancement method using a model of mutual inhibition for noise reduction in speech which is contaminated by white noise or background noise mentioned above. For evaluating speech quality we used the spectral distortion ratio (SD).

キーワード 雑音処理、背景雑音、相互抑制、音声強化

1 まえがき

近年、音声認識装置の性能が向上し、実用化が広く進んで来ているが、種々の雑音により装置の信頼性が大幅に低下することが問題となっている。特に、定常な背景雑音のほかに、道路での自動車の走行雑音やプリンタの駆動雑音などの非定常な多様な雑音が問題となっている。これらの非定常な雑音の中には、そのスペクトルの形状が音声や音声のピッチ成分に類似しているものも多く、音声の認識率の著しい低下を招く原因となっている [1]。

しかしながら雑音の除去や軽減法に関しては線形なウィナーフィルタなどによるものの他に、聴覚の末梢系を相互抑制機構が働く適応フィルタと考え、それを音声の雑音除去に応用しようとする森田らの先駆的研究がある [2]。彼らは入力音声を24チャンネルよりなる二つのフィルタバンクに分け、一方のバンク出力は整流し対数変換を施した後、バンク間の相互抑制を行ない、振幅の大きいフォルマント部を強調する。この出力信号でもう一方のバンク出力の振幅を調整し、相対的に雑音に対し信号成分を強調し、その出力を合成することにより特性の改善された音声を得ている。しかし、改善率の定量的評価は行なっていない。

最近のこの種の研究としては Cheng ら [3]、Dang ら [4] および Shamma [5] の研究がある。特に Cheng らはヘリコプター雑音に対して相互抑制が有効であることを述べており、SN 比に関しても理論的な解析を行ない、また実験的にもその有効性を証明している。彼らは森田らと異なり音声をサンプリング後離散ケプストラム変換を施し、その出力に対して相互抑制をかけている。

本研究は聴覚機構を生理学的に忠実に模擬するのではなく、工学的に相互抑制を応用しようとする立場に立ち、Cheng らの音声強化システムに振幅調整係数を導入し、(1) 雑音の種類、(2) 相互抑制の周波数帯域幅、(3) サンプル数 などの変化に対する SD (スペクトルひずみ尺度) を求め、音声の特性

の改善に必要な基礎資料を得ている。Cheng らがその評価に SN 比を用いているのに対し、本研究では音声の明瞭度に関係の深い SD を用いた。

本論文では、白色雑音ばかりでなく現実に存在する道路での自動車の走行雑音やプリンタの駆動雑音のような有色の背景雑音によって劣化された音声をも対象としている。以下本研究で用いられた聴覚生理学に基づく音声強化アルゴリズムを示す。このアルゴリズムは種々の雑音環境下において雑音レベルの減少に有効であり、かつ音声品質の改善を行ない、より明瞭度の高い音声再生できることを実験的に明らかにしている。

2 音声強化システム

図 1: 音声強化システム

Fig.1 Speech enhancement system

本論文に用いた音声強化システムを図 1 に示す。まず、8kHz でサンプリングされた雑音で汚された音声信号を 128 点からなるハミング窓 $W_1(t)$ を通して、ケプストラム変換

(Cepstral Transform) をする (上の経路)。得られたケブストラムから方形窓 $W_2(t)$ によって短時間領域成分のみを取り出し、フーリエ変換 (FFT) によって音声信号のスペクトル包絡を得る。フレーム単位の遅延 (D_i) を介して、フレーム単位で重み付け (W_i) されたスペクトルの平均 (Spectral Average) を取る。その際、3 フレーム分の遅延が起こる。次に、そのスペクトル成分を周波数空間において相互抑制 (Function of Spatial Lateral Inhibition, FSLI) する。

相互抑制によって生じた負の成分は整流器 (Rectifier) によって除去し、それを振幅成分とする。一方、別の経路 (下の経路) でフーリエ変換されて 3 フレーム分遅延されて取り出された信号を位相成分と振幅成分とに分離し、振幅成分を (入力の振幅成分 $\times (1 + R \times$ 整流器出力)) によって変換し、位相成分と合成したものを逆フーリエ変換 (IFFT) することによって強化された音声信号を再生する。

3 音声の特性改善法

3.1 雑音モデル

音声信号を $s(t)$ とし、雑音の加わった音声信号を、

$$x(t) = s(t) + k \times n(t) \tag{1}$$

で表されるものとする。ただし、 $n(t)$ は雑音信号、 k は雑音強度を表す係数である。このフーリエ変換を取ると

$$X(e^{j\omega}) = S(e^{j\omega}) + k \times N(e^{j\omega}) \tag{2}$$

となる。

3.2 ケブストラム分析

図 2: ケブストラム分析の手順
Fig.2 Arrangement of cepstrum analysis

音声波形 $x(t)$ は、音源波形 (ピッチ波形) $g(t)$ と声道のインパルス応答 $v(t)$ との畳み込みとして表現される (放射の影響は考えないものとする) [6]。すなわち

$$x(t) = g(t) * v(t) \tag{3}$$

離散フーリエ変換の絶対値の対数演算操作を D とすると

$$\begin{aligned} D\{x(t)\} &= \log|X(e^{j\omega})| = D\{g(t) * v(t)\} \\ &= D\{g(t)\} + D\{v(t)\} \end{aligned} \tag{4}$$

$D\{x(t)\}$ の逆フーリエ変換をケブストラム (cepstrum) $c(t)$ と呼ぶ。すなわち

$$\begin{aligned} c(t) &= \frac{1}{2\pi} \int_0^{2\pi} \log|X(\omega)| e^{j\omega t} d\omega \\ X(\omega) &= X(z)|_{z=e^{j\omega}} \end{aligned} \tag{5}$$

よって、 $x(t)$ のケブストラムは $g(t)$ のケブストラムと $v(t)$ のケブストラムの和となる。ケブストラムの独立変数は時間の次元 (quefrency) をもつ。

有声音の場合には、 $D\{g(t)\}$ のケブストラムは時間軸上で $1/F_0$ (F_0 : ピッチ周波数) の近傍における成分として、また $D\{v(t)\}$ のケブストラムは短時間領域成分として現れる。ゆえに、ケブストラムに窓をかけて短時間領域成分のみを取り出し ($g(t)$ の除去)、これに離散的フーリエ変換を施せば、スペクトル包絡が得られる。ケブストラム分析の手順を図 2 に示す。すなわち図 1 の音声強化システムのコンボリューション入力として平均化された音声信号のスペクトル包絡が得られることになる。

3.3 スペクトルの平均

音声信号のスペクトルはフレーム間で急変しないが、雑音はフレーム間で不規則なスペクトル変動を引き起こすと考えられる。そこで、フレーム間の雑音による不規則なピークを減少させる一つの方法としてスペクトルの平均を取る。重み付けされたスペクトルの平均を

$$\bar{P}_x^{(i)}(\omega) = \frac{1}{2N+1} \sum_{j=-N}^N W_j P_x^{(i-j)}(\omega) \tag{6}$$

とした。

本実験においては $N=2$ とし、荷重として $[W_0, W_1, W_2, W_3, W_4] = [0.7, 1.1, 1.4, 1.1, 0.7]$ とした。 $\bar{P}_x^{(i)}(\omega)$ は平均化された (i) 番目のフレームの短時間電力スペクトルである。

3.4 相互抑制

図 3: 相互抑制関数のインパルスモデル
Fig.3 Impulse model of a lateral inhibition function

FSLI は内耳の基底膜における神経相互間の相互抑制機構を模擬したものであり、音声のスペクトルの山の部分をシャープにすることと谷の部分の雑音を軽減することから音声強化に有効であると考えられる。相互抑制の一般的機能は空間的入力パターンあるいは時間的な入力変動をシャープにすることである [3] [4] [5] [8]。

図 3 は本実験に用いた FSLI の特性を示したものであって ($B_f = 5$ の場合)、横軸は周波数標点を示し、縦軸は周波数 $B = 0$ の位置に入力 1 が加えられた場合のその近傍の標

本点に得られる出力 (インパルス応答) を示している。なお、 B_f は FSLI の広がり を定めるパラメータである。

FSLI のパラメータは

$$P_l + P_c + P_r = 0 \quad (7)$$

のように設定する。この制限は相互抑制により雑音の和の平均値が零となって雑音の軽減が計られるためである。本実験では以下のような値を用いた。

$$P_c = 1 \text{ and } P_l = P_r = -0.5 \quad (8)$$

相互抑制された出力は $\bar{P}_x^{(i)}(\omega)$ と図 3 に示す相互抑制関数とのコンボリューションによって得られる。

図 4 は相互抑制を受けた出力 (— 印) とその入力 (⋯ 印) とを示している。入力 は “aioi” を 8kHz でサンプリングしたときの音声信号の/a/の部分に白色雑音を加わったものであり、これに対するスペクトル平均を示し、スペクトルの大きな山の部分がホルマントである。相互抑制効果により、そのホルマント部が強調されている様子がわかる。

図 4: 白色雑音の加わった音声信号に対するのスペクトル平均 (点線) と相互抑制によるホルマントの強調 (実線)

Fig.4 Spectrum average(dotted line) and formant enhancement by lateral inhibition(solid line) for speech signal with white noise

4 実験条件と評価法

4.1 実験条件

以下に述べるような条件の下に実験を行なった。

1. 音声データはすべて男性話者によるものを使用する。例えば、音声データとして “aioi”、“bizen”、“hachioji”、“shimonoseki”、“kisarazu”、“sendai” などを用いた。
2. 各音声データは最長 3 秒間である。
3. 背景雑音としての白色雑音はコンピュータによって作成した疑似ガウス白色雑音、自動車の走行雑音は交通量の多い道路で録音し、プリンタの駆動雑音はプリンタ動作時に録音したものをサンプリング周波数 8kHz で A/D 変換したものを使用する。

4. 音声および雑音の周波数帯域は 0 ~ 3 kHz とする。

5. 音声処理は 1 フレーム 256 サンプルとしハミング窓によって重ならないように切り出す。

4.2 実験結果の評価法

再生信号の評価は音声明瞭度と関係の深いスペクトルひずみ尺度 (SD) [7] を用いた。

$$SD = \sqrt{\frac{1}{N_F} \sum_{i=1}^{N_F} \frac{1}{W} \int_0^W \{S_s^{(i)}(f) - S_y^{(i)}(f)\}^2 df} \quad (dB) \quad (9)$$

ここで、 N_F は測定区間のフレーム数 (本実験においては $N_F = 30 \sim 80$)、 W は信号の帯域幅、 $S_s^{(i)}(f)$ および $S_y^{(i)}(f)$ は入力信号 $s(t)$ と出力信号 $y(t)$ の対数スペクトル (dB) であり、次式で定義される。

$$\begin{aligned} S_s^{(i)}(f) &= 10 \log_{10} |S(f)|^2 \\ S_y^{(i)}(f) &= 10 \log_{10} |Y(f)|^2 \end{aligned} \quad (10)$$

ここで、 $S(f)$ 、 $Y(f)$ はそれぞれ周波数 f における入力信号と出力信号のスペクトルである。

5 実験結果および考察

以上に述べたような基本的な構成条件の下で、白色雑音、自動車の走行雑音、プリンタの駆動雑音などが加わった音声入力に対し、振幅成分調整係数 R を変えた場合の FSLI の効果を SD を評価基準として求めた。

5.1 スペクトルの振幅成分調整係数 R に対する効果

表 1 と表 2 はそれぞれ白色雑音の場合について母音の多い単語と子音の多い単語に対して、雑音係数 k と振幅成分調整係数 R とをパラメータとしたときの時間信号 $x(t)$ の入力 SN 比と出力 SD とを求めたものである。

表 1 と表 2 のスペクトルひずみ尺度 (SD) の評価を見ると、スペクトル調整係数 R を調整することにより、それぞれの雑音強度 k に対して最適な R の値 (dB) が存在することがわかる (太字で強調)。例えば、表 2 の $k=10$ の場合、最適な R の値に対して SD が 13.01dB である。これは $R=0$ のときの雑音を含んだ原音に対する SD である 20.42dB より 7.41dB の改善が見られる。SD の絶対値から異なっているので単純に比較は出来ないが、電話の PCM 方式の評価データによると、20dB の SN 比での改善量が SD では約 3.5dB に対応することを考えると (文献 [7] の Fig.3)、SD における 7.41dB の改善量は相当有効であると言えよう。なお、表中 S/N は $x(t)$ の信号対雑音比である。実際に、試聴した結果 SD が小さいほど良好な再生音声を得られた。表 2 の子音の多い単語に対する結果より、子音が多いと、このシステムでは効果が低くなることがわかる (SD が大きくなる)。

表 3 と表 4 は FSLI の効果を示すためのものであって、“処理なし” は $x(t)$ の SD を示し、“FSLI なし” は図 1 において Convolution with FSLI と Rectifier を除き、 $R=1$ とした場合の SD を示している。

FSLI を用いた場合の方が母音の多い単語と子音の多い単語のいずれの場合についても SD が低く FSLI が有効であることがわかる。

表 1 振幅成分調整係数 R の効果 (“aioi”、 $B_f = 8$)
Table 1. The effect of the adjusting coefficient R
for amplitude component(in the case of
“aioi” and $B_f = 8$)

雑音強度 k	S/N (dB)	振幅成分調整係数 R				
		0.0	1.0	2.0	3.0	4.0
0	∞	0.00	12.33	16.14	18.46	20.12
2	4.17	13.45	9.33	10.28	11.47	12.54
4	-1.85	15.76	10.78	10.17	10.50	11.06
6	-5.38	16.66	11.87	10.72	10.65	10.93
8	-7.87	17.10	12.58	11.22	10.95	10.08
10	-9.81	17.33	12.97	11.53	11.19	11.26

表 2 振幅成分調整係数 R の効果 (“hachioji”、 $B_f = 8$)
Table 2. The effect of the adjusting coefficient R
for amplitude component(in the case of
“hachioji” and $B_f = 8$)

雑音強度 k	S/N (dB)	振幅成分調整係数 R				
		0.0	1.0	2.0	3.0	4.0
0	∞	0.00	13.37	17.46	19.94	21.71
2	0.89	16.19	12.49	13.14	14.06	14.94
4	-5.13	18.38	13.07	12.61	12.94	13.46
6	-8.65	19.45	13.79	12.78	12.76	13.04
8	-11.15	20.04	14.36	13.07	12.84	12.98
10	-13.09	20.42	14.80	13.35	13.01	13.06

表 3 FSLI の効果を示す SD (“aioi”、 $B_f = 8$)
Table 3. SD effect of FSLI(in the case of
“aioi” and $B_f = 8$)

雑音強度 k	処理なし	FSLI なし	FSLI あり
0	0.00	5.92	0.00
2	13.46	12.25	9.33
4	15.78	15.22	10.17
6	16.68	16.38	10.65
8	17.12	16.93	10.95
10	17.35	17.21	11.19

表 4: FSLI の効果を示す SD (“hachioji”、 $B_f = 8$)
Table 4. SD effect of FSLI(in the case of
“hachioji” and $B_f = 8$)

雑音強度 k	処理なし	FSLI なし	FSLI あり
0	0.00	9.95	0.00
2	16.44	15.16	12.49
4	18.61	17.72	12.61
6	19.67	19.03	12.76
8	20.25	19.78	12.84
10	20.63	20.26	13.01

5.2 各種雑音と相互抑制幅 B_f に対する FSLI の効果

図 5 は 子音の多い単語に対して、FSLI を用いた場合の白色雑音、プリンタの駆動雑音、自動車の走行雑音の大きさに
対して最適な係数 R を選んだ場合の SD の変化をプロット
したものであり、記述の順番に SD が改善されている。

図 6 と図 7 は 子音の多い単語に対して、FSLI を用いた場
合と FSLI を用いなかった場合のプリンタの駆動雑音と自動
車の走行雑音に対する SD を比較したものである。両方の雑
音に対し、FSLI を用いた場合の方が FSLI を用いなかった
場合より SD がほぼ 3dB 改善されている。しかし、プリン
タの駆動雑音の方が自動車の走行雑音より SD の絶対値が約
2dB 小さい。以上の結果より、FSLI を用いることがいづれ
の場合に対しても有効であることがわかる。

図 8 は 子音の多い単語に対して、FSLI の係数 B_f を調整
したときのプリンタの駆動雑音に対する SD を比較したもの
である。FSLI の係数の幅 B_f が $B_f = 8$ 、 $B_f = 11$ 、 $B_f = 5$
の順番に SD が改善されている。実験結果より、 $B_f = 8$ の
場合が最も有効であることがわかる。

図 9 は 子音の多い単語に対して、フレーム中のサンプル
数を 32、64、128、256、512、1024 に変えたときの雑音強
度 $k=2$ に対する、白色雑音、プリンタの駆動雑音、自動車の
走行雑音に対する SD を比較したものである。図より、白
色雑音、プリンタの駆動雑音、自動車の走行雑音の順番に
SD が改善されていることがわかる。サンプル数が増加する
につれて SD は大きくなる傾向にあり、64 サンプルのとき
が最も有効であり、サンプル数が多ければ多いほど良いとい
うものではないことがわかる。

また、図 10、11、12 はサンプル数を 128 にしたとき
に “aioi” を 8kHz でサンプリングしたときの音声信号の /a/
の部分を示したものであり、白色雑音のないときの入力音声
波形と白色雑音を加わった入力音声波形と再生音声波形とを
それぞれ示したものであり、雑音除去の様子がわかる。

図 5: 各種雑音に対する FSLI を用いた場合の効果の比較
 (“hachioji”、 $B_f = 8$ の場合)

Fig.5 Comparison between the effects of FSLI for various
noises(in the case of “hachioji” and $B_f = 8$)

図 6: プリンタの駆動雑音に対する FSLI を用いた場合と FSLI を用いなかった場合の比較 (“hachioji”、 $B_f=8$ の場合)

Fig.6 Comparison between the cases with FSLI and without FSLI for printer noise (in the case of “hachioji” and $B_f=8$)

図 7: 自動車の走行雑音に対する FSLI を用いた場合と FSLI を用いなかった場合の比較 (“hachioji”、 $B_f=8$ の場合)

Fig.7 Comparison between the cases with FSLI and without FSLI for road noise of automobile (in the case of “hachioji” and $B_f=8$)

図 8: プリンタの駆動雑音に対する FSLI の係数 B_f を調整したときの比較 (“hachioji”、 $B_f=8$ の場合)

Fig.8 Comparison between the cases when coefficient B_f of FSLI is adjusted to four values in the case of printer noise(in the case of “hachioji” and $B_f=8$)

図 9: フレーム中のサンプル数を変えたときの SD の変化 (“hachioji”、 $B_f=8$ 、 $k=2$ の場合)

Fig.9 The variation when sampling numbers in a frame are changed(in the case of “hachioji” and $B_f=8$ and $k=2$)

図 10: 白色雑音に加わっていない入力音声信号の波形 ($k=0$ の場合)

Fig.10 Input of clean speech signal(in the case of $k=0$)

図 11: 白色雑音に加わった場合の入力音声信号の波形 ($k=3$ の場合)

Fig.11 Input of contaminated speech signal with white noise(in the case of $k=3$)

図 12: 再生音声信号の波形 ($k=3$ の場合)
Fig.12 Waveform of output speech signal
(in the case of $k=3$)

6 むすび

人間の聴覚システムにおける相互抑制機構が雑音抑制効果があることに着目し、それを工学的に応用する一つの試みとして、本システムを提案し、それがスペクトルひずみ尺度の下で有効であることを種々の雑音について実験的に検証した。

実験結果をまとめると、次のような結論が得られる。

1. それぞれの雑音強度 k に対して最適なスペクトル調整係数 R が存在する。実際に、視聴した結果においても SD が小さいほど良好な再生音声を得られる。また、処理なしの音声は SD が低くても耳ざわりな音として聞こえる。これは聴覚のマスキング効果によるものと思われる。
2. 母音の多い音声の方が子音の多い音声より雑音軽減の効果が大きい。
3. プリンタの駆動雑音、自動車の走行雑音など、雑音の種類によって $FSLI$ の雑音軽減効果が異なる。
特に、白色雑音に対する $FSLI$ の雑音軽減効果が顕著である。
4. 相互抑制の係数 B_f には最適値が存在する。
5. 1 フレーム当たりのサンプリング数にも最適値が存在する。

以上、音声信号の雑音軽減のために $FSLI$ が特に白色雑音に対して有効であることを実験的に確かめたが、今後の研究課題としては自動車の走行雑音などのような有色雑音によって劣化された音声に対しても、さらに強化する方法の検討が挙げられる。

また、本研究においては SD を評価基準として用いたが、音声の良さの評価は音声明瞭度が最善であり、明瞭度の下での音声強化の方法を検討して行きたいと考えている。

本研究は相互抑制機構のほんの基本的な特性と応用の可能性を示したものである。今後さらに検討を深めることによってさらに有効な応用法が開けるものと考えている。

(平成 7 年 1 月 30 日受付、同 7 年 3 月 24 日再受付)

参考文献

- [1] 石田 明, 小畑 秀文: "実環境下での音声/非音声の判別", 日本音響学会誌 47 巻 12 号, pp.911-917 (1991).
- [2] 森田 龍弥, 名倉 和宏, 藤井 克彦: "マルチチャネルフィルタによる雑音処理機構", 昭和 57 年度電子通信学会総合全国大会, S7-9 (1982).
- [3] Yan Ming Cheng and Douglas O'Shaughnessy: "Speech enhancement based conceptually on auditory evidence", IEEE Trans. on Signal Processing, Vol.39, No.9 (1991).
- [4] V.C. Dang, R. Carre, D. Tuffelli: "Speech signal preprocessing taking into account lateral inhibition" Signal Processing III: Theories and Applications., vol. 1, pp. 533-536 (1986).
- [5] Shihab A. Shamma, "Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve" J. Acoust. Soc. Amer., pp. 1622-1632 (1985).
- [6] Steven. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech Signal Processing, vol. ASSP-27, no.2, pp. 113-120 (1979).
- [7] 伊藤 憲三: "音声のデジタル波形符号化方式の客観的品質評価尺度の検討", 電子通信学会論文誌 Vol.J 66-A No.3, pp.274 - 281 (1983).
- [8] 目片 強司, 山田 義則, 鈴木 良二: "補聴器への応用を目的としたホルマント強調の一検討- 明瞭度を尺度とした評価 -, " 信学技報 SP 93 - 35 ~ 43, vol.93, no.155, pp. 1-8 (1993).

崔 載 承

(非会員) 1962 年 10 月 29 日生まれ。
89 年 2 月 韓国 朝鮮大学工学部電子工学科卒業。95 年 3 月 大阪市立大学工学部電気工学科修士課程修了。主として雑音中の音声信号処理に関する研究に従事。電子情報通信学会会員。

岡本 次郎

(正員) 1964 年 大阪市立大学大学院電気工学科修士課程修了。同年同大学電気工学科助手。70 年同科講師。88 年同科助教授。工学博士。その間音声情報処理、2 次元デジタルフィルタの安定性、ニューラルネットワーク、樹木の成長のモデル化などの研究に従事。

細川 省一

(正員) 1960 年 大阪市立大学大学院修士課程修了。同年 5 月同大学工学部助手、63 年同講師、71 年同助教授、90 年同教授 (情報処理工学講座)、現在に至る。工学博士。この間、音声信号の高効率符号化法および伝送、コンピュータネットワーク、ニューラルネットワークの研究に従事。