音声言語情報処理 39－11
(2001. 12. 20)

音声のピッチ変動の耐雑音音声認識における効果

相川 清明† 石塚 健太郎†

† 日本電信電話（株），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: †aik@idea.brl.ntt.co.jp, ††ishizuka@atom.brl.ntt.co.jp

あらまし 分析フレーム内での繰り返し波形を分割、加算平均した1周期の波形を用いることにより耐雑音性を高めた新しいスペクトル推定法 PHASOR を提案する。提案アルゴリズムは雑音下母音知覚実験結果に基づいている。音声スペクトルは、従来のようにフレーム内の全波形から求めるのではなく、繰り返しの単位波形から求める。繰り返しの単位波形は、声門の開閉タイミングの時間的揺らぎを補正して求める。これにより、周期性雑音、ランダム雑音の両方に対して耐雑音性が向上する。不特定話者、特定話者両音声認識実験により、PHASOR の耐雑音性効果を示した。本方法は有声音の認識誤りを大幅に減少できる。PHASOR により、従来のスペクトル減算法やケプストラム平均正規化法と比べて高い認識性能が得られた。

キーワード 耐雑音性，音声認識，ピッチ，同期，スペクトル分析

An Effect of Pitch Fluctuation on Noise-Robust Speech Recognition

Kiyoaki AIKAWA† and Kentaro ISHIZUKA†

† NTT Communication Science Laboratories, NTT Corporation
3–1 Morinosato-Wakamiya, Atsugi-Shi, Kanagawa, 243–0198 Japan

E-mail: †aik@idea.brl.ntt.co.jp, ††ishizuka@atom.brl.ntt.co.jp

Abstract This report proposes a new noise-robust spectral estimation method called PHASOR. The new method is motivated by the effect of pitch fluctuation on noisy speech perception. The PHASOR estimates speech spectrum from a single pitch period of speech signal obtained by summing multiple pitch periods in a frame, whereas conventional method uses the whole signal in the frame. PHASOR suppresses the effect of additive noises. Speaker-dependent and speaker-independent phoneme recognition experiments demonstrate that the PHASOR greatly reduces the phoneme recognition error rate for noisy speech data. The PHASOR also outperforms conventional noise reduction methods, cepstral mean normalization and spectral subtraction.

Key words noise-robust, speech recognition, pitch, synchronous, spectral analysis

1. はじめに

音声認識では通常 30ms 程度の時間窓で切り出された音声波形からスペクトルを推定している。この単位は通常フレームと呼ばれている。フレーム長は安定にスペクトルを求められるように有声部に見られる波形の周期の 2 倍以上で、ピッチやスペクトルが一定とみなせる長さが用いられてきた。従来法ではフレームに入る音声信号を単一の時系列として扱いスペクトルを求める。この従来法をフレーム単位の方法と呼ぶことにする。

ゆっくりとした読み上げ音声では 30ms というフレーム長は音素長よりも短く、スペクトルもピッチもある程度一定とみなせるが、自然発声の音声では必ずしもそうではない。それだけではなく、自然な音声には図 1に模式的に描いたようなピッチの揺らぎが存在する。すなわち、繰り返し波形の間隔が不均一になる。従来のフレーム単位の方法では、ピッチの揺らぎがあると、各波形で位相がずれたことになり、信号のエネルギーが低下する。

最近、著者らは雑音下の母音知覚について興味ある知見を得た。まず、同じスペクトル包絡の背景雑音でも、調波性雑音の方がランダム雑音に比べて母音知覚実験における正答率が高い [1]。また、母音の基本周波数に揺らぎがある方が、特に周期性のある雑音下での母音正答率が高い [2], [3]。この結果は音声の揺らぎを積極的に用いることにより雑音を抑制する方法の可能性を示唆している。もし、1 フレーム内で繰り返された波形を位相を合わせて加算平均できれば、位相がまちまちのままスペクトルを求めた場合に比べ、音声信号エネルギーの減衰を防げる可能性がある。この操作は、周期的な雑音が音声に重畳されていたとしても、雑音の位相をずらして加算することになるので、雑音を抑圧ができる可能性がある。本報告ではこのフレーム内波形処理を用いた耐雑音スペクトル推定法 PHASOR (PHase Adjusted Sum Of Responses) を提案する。

PHASOR では 1 フレームより短い分析窓を使うことになる。1 フレームよりはるかに短い分析窓で分析したスペクトルを統合して 1 フレームのスペクトルを推定する試みは提案されている [4]。この方法は、1 フレームの中のエネルギーの強い部分のスペクトルを強調して加算することにより、耐雑音性を高める方法である。しかしフレームの内部での波形処理による雑音抑制は考えられていなかった。

本報告では、まず、PHASOR のアルゴリズムを述べ、次に PHASOR の雑音抑制効果について述べる。実験では、特定話者及び不特定話者音素認識実験により PHASOR の耐雑音性を評価する。また、従来の雑音抑制方法であるスペクトル減算法 SS (Spectral Subtraction) [5] および、ケプストラム平均正規化法 CMN (Cepstral Mean Normalization) [6] との比較を行なう。

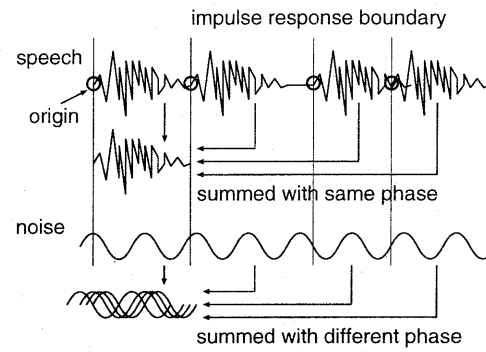

図 1　音声の繰り返し波形の位置を合わせて加算することにより、重畳した雑音が抑制される。雑音は非同期的に加算されるためである。(周期性雑音の場合)

2. PHASOR のアルゴリズム

PHASOR により 1 フレームのスペクトルを求める方法を述べる。音声の有声部分の波形には図 1に模式的に描いたような周期構造が見られる。このような周期波形は、声門の周期的開閉により形成される。声門の開閉はほぼ周期的であるが、時間的な揺らぎがある。これらの繰り返し波形を時間揺らぎを除去して加算平均すれば、フレームの中での繰り返し単位の波形を精度良く求めることができる。この波形は、声道のインパルスレスポンスのうち、1 ピッチ周期を越える部分が波形の前の部分に重畳されたもとの考えることができる。次の繰り返し波形に重畳する部分の始点はピッチパルス間隔の揺らぎの影響を受ける。しかし、声道インパルスの 1 ピッチを越える分はエネルギーが減衰していること、影響するものはせいぜい隣接した繰り返し波形であることなどから、揺らぎの影響は大きくはないと考えられる。

PHASOR では、フレーム内で逐次的にピッチ周期を求める。始めに 1 フレーム内での分析の開始時間 k_1 をゼロ、すなわちフレーム内の時間原点に設定する。音声信号を $s(k)$ と記述すると、隣接する 2 小区

間の相互相関は、

$$c(k_l, n) = \frac{d_{12}(k_l, n)}{\sqrt{d_1(k_l, n)\, d_2(k_l, n)}} \quad (1)$$

$$d_{12}(k_l, n) = \sum_{k=0}^{n-1} s(k_l + k)s(k_l + n + k) \quad (2)$$

$$d_1(k_l, n) = \sum_{k=0}^{n-1} s(k_l + k)s(k_l + k) \quad (3)$$

$$d_2(k_l, n) = \sum_{k=0}^{n-1} s(k_l + n + k)s(k_l + n + k). \quad (4)$$

のように与えられる。ここで、k は時間に対応する。この式は、k_l から始まる長さ n の部分波形と $k_l + n$ から始まる長さ n の部分波形の相互相関を求めている。ピッチ周期は、相互相関の最大値をあたえる n として求まる。

$$k_{max} = \underset{n_{min} \leqq n \leqq n_{max}}{\text{argmax}} c(k_l, n). \quad (5)$$

ここで、サンプル数で表されるピッチ周期の最小値 n_{min} と最大値 n_{max} は基本周波数の最小値 f_{min}、最大値 f_{max}、および、サンプリング周波数 f_s から

$$n_{min} = \frac{f_s}{f_{max}} \quad (6)$$

$$n_{max} = \frac{f_s}{f_{min}}. \quad (7)$$

のように求まる。なお、本報告では $f_{min} = 80$ Hz、$f_{max} = 400$ Hz とした。以上により、ピッチ周期 k_{max} が求まったら、

$$k_{l+1} = k_l + k_{max}. \quad (8)$$

により、始点をずらして相互相関を求める操作を繰り返す。繰り返しの単位となる波形は、始点と継続時間の組として求まる。これらの波形を位置ずれを補正して加算する。

実際の演算では、波形が得られるたびに波形蓄積バッファ $r_l(k)$ に足し込んでいく。各波形は周期的な波形の 1 周期分であるが、必ずしも始点が対応しているとは限らない。特に、途中に周期性が乱れる部分があれば、位置ずれが生じる可能性が高い。m 番目の波形の時間補正の値を j_m とすると、l 個までの波形の和は

$$r_l(k) = \sum_{m=1}^{l} s(k_m + j_m + k) \quad (9)$$

により与えられる。

$(l+1)$ 番目の波形をそれまでに得られた波形の和に適合させるための時間ずらし幅を j_{l+1} とする。j_{l+1} はそれまでに加算された波形 $r_l(k)$ と $(l+1)$ 番目の波形 $s(k_{l+1} + j + k)$ の相互相関 $q(j)$ を最大化する値として、

$$j_{l+1} = \underset{j_{min} \leqq j \leqq j_{max}}{\text{argmax}} q(j) \quad (10)$$

のように与えられる。ここで、

$$q(j) = \frac{g_{12}(j)}{\sqrt{g_1(j)g_2(j)}} \quad (11)$$

$$g_{12}(j) = \sum_{k=0}^{K} r_l(k)s(k_{l+1} + j + k) \quad (12)$$

$$g_1(j) = \sum_{k=0}^{K} r_l(k)r_l(k) \quad (13)$$

$$g_2(j) = \sum_{k=0}^{K} s(k_{l+1} + j + k)s(k_{l+1} + j + k). \quad (14)$$

である。

$(l+1)$ 番目の波形の和は $(l+1)$ 番目の波形を波形バッファに

$$r_{l+1}(k) = r_l(k) + s(k_{l+1} + j_{l+1} + k) \quad (15)$$

のように逐次的に足し込むことにより求まる。図 2 は以上の演算のフローチャートを示す。図中で "impulse response" は声道インパルスレスポンスのことを意味し、1 周期の波形に対応する。argmax のブロックではループによる最大値探索の演算が行なわれる。これにより得られた波形を足し込んだ波形の数で割れば、フレーム内の繰り返し波形の 1 周期に相当する波形が求まる。求められた 1 周期の波形から、自己相関関数を求め、LPC 分析を行なえば、スペクトルやケプストラム係数を求めることができる。ケプストラム係数から、デルタケプストラムを求めることもできるし、CMN もケプストラム係数に対して行なうことができる。SS は一旦スペクトルを求めて行なう必要がある。

次に、無声区間での処理方法について述べる。無声区間、無音区間ではフレーム内で周期性は見られない。そこで、フレームを等分して波形を加算するという方法が考えられる。しかし、無声区間でも有声音が重畳していることがあり、ある程度の周期性が見られる場合がある。そこで、本報告では、無声区間であっても有声区間と同一の処理を行なった。

3. 雑音抑制効果

PHASOR による雑音抑制効果を述べる。PHA-

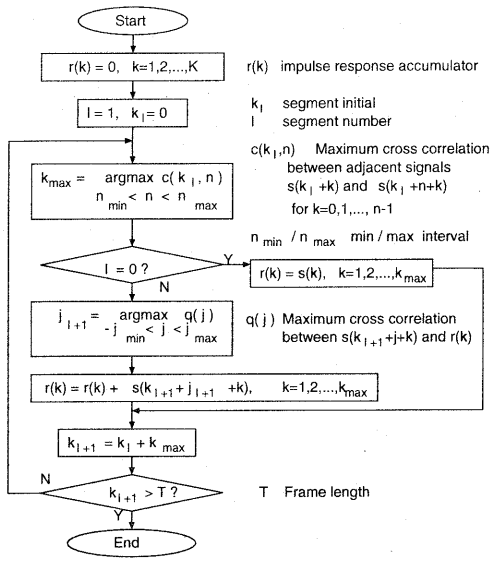

図2 フレームデータから繰り返しの単位波形を求める演算のフローチャート

SORは図1のように、繰り返し波形を分割し位相を同期させて加算平均する。声門からのパルス間隔には揺らぎがある。すなわち、繰り返し波形の位相には揺らぎがある。音声信号を同期加算平均すると、音声と位相が同期した成分以外の雑音は周期的な雑音であっても異なる位相で加算されるので抑制される。

まず、信号がPHASORの処理によりどのように求められるかを示す。音声波形は位相を同期させて加算される。1ピッチ周期の信号を $s(k)$、繰り返し数を I とすると、加算平均した音声信号のエネルギーは

$$P_s = \frac{1}{L}\sum_{k=1}^{L}(\frac{Is(k)}{I})^2 \tag{16}$$

$$= \lambda^2 \tag{17}$$

となる。ここで、λ^2 は繰り返し波形1周期分の音声のエネルギーである。また、L はピッチ周期である。

次に、雑音のエネルギーを求める。もし、ピッチの揺らぎがランダムな場合、正弦波状の雑音はランダムに位相をずらして加算されることになる。加算平均された波形は

$$N(k) = \frac{1}{I}\sum_{m=1}^{I}\sin(\omega k\Delta t + \phi_m) \tag{18}$$

$$= \frac{1}{I}\sum_{m=1}^{I}(\sin(\omega k\Delta t)\cos(\phi_m) \tag{19}$$

$$+ \sin(\phi_m)\cos(\omega k\Delta t)) \tag{20}$$

で表される。ここで、ω は正弦波状雑音の周波数、Δt はサンプリング間隔を表す。もし、位相 ϕ_m が $-\pi < \phi_m \leqq \pi$ の範囲でランダムな値をとるならば、$\sin(\phi_m)$ と $\cos(\phi_m)$ は近似的にランダムな変数 v_m と u_m に置き換えられる。従って、式(20)は

$$N(k) = \frac{1}{I}\sum_{m=1}^{I}(u_m\sin(\omega k\Delta t) \tag{21}$$

$$+ v_m\cos(\omega k\Delta t)) \tag{22}$$

と、近似できる。繰り返し1周期分の雑音のエネルギーを ρ^2 とすると、フレーム内でのI個の繰り返しでは、総エネルギーは

$$P_n = \frac{1}{L}\sum_{k=1}^{L}(N(k))^2 \tag{23}$$

$$\cong \frac{\rho^2}{I}. \tag{24}$$

となる。

もし、周期的雑音でなくランダム雑音 $x_m(k)$ の場合には、雑音波形の加算平均は

$$N(k) = \frac{1}{I}\sum_{m=1}^{I}x_m(k) \tag{25}$$

となる。従って、フレームでの平均雑音エネルギーは

$$P_n = \frac{1}{L}\sum_{k=1}^{L}(\frac{1}{I}\sum_{m=1}^{I}x_m(k))^2 \tag{26}$$

$$= \frac{\sigma^2}{I} \tag{27}$$

となる。ここで、σ^2 は長さ L の音声の1周期における雑音のエネルギーである。

従って、有声部分においては重畳する雑音がランダムな場合にはPHASORによりSNR（信号対雑音比）は I だけ改善される。また、雑音が周期的な場合には、音声のピッチの揺らぎに依存して最大 I だけSNRが改善される。無声部分で音声信号に全く周期性が無い場合にはSNRは従来のフレーム単位の方法と同じになる。

4. 実　験

日本語23音素（18子音と5母音）について特定話者及び不特定話者音素認識実験を行なった。音声データのサンプリング周波数は12 kHzである。音素モデルには3状態のコンテキスト独立のHMMを用いた。出力確率は8混合ガウス分布で表される。

実効的なフレーム長は35 ms、フレームシフトは10 msである。実験ではPHASORと従来のフレーム単位の方法で認識率を比較した。用いた音声は、

雑音の無いものと、SNR 20 dB でピンク雑音を重畳したものである。音声認識のための特徴パラメータとしては、ケプストラムとデルタケプストラムを用いる。

PHASOR と従来のフレーム単位の方法の違いは、LPC 分析に用いる波形である。PHASOR では、加算平均した繰り返し波形 1 周期分を用い、従来法では、1 フレームの波形を用いる。

4.1 特定話者音素認識

特定話者音素認識は ATR の日本語重要語 5240 単語データベース男性 1 名分を用いて行なわれた。データを 2620 単語ずつに 2 分し、その片方を用いて HMM を学習し、もう 1 方を用いて認識実験を行なった。

図 3 に 23 音素認識結果を図 4 に 5 母音認識結果を示す。これらの図では、従来法による雑音がない場合の結果を参考結果として示し、雑音を付加した音声の認識結果を LPCCEP で記述した従来法と PHASOR で記述した提案法を比較している。デルタケプストラムを用いない場合と用いる場合を組として棒グラフを示す。これらの図から、PHASOR により雑音下での音声認識率を向上できることがわかる。特にデルタケプストラムを併用しない 5 母音認識の場合には、PHASOR により認識誤りを半分近くに減少できる。また、23 音素認識ではデルタケプストラムを併用することによりさらに認識率を向上できる。

図 5 は従来法（LPCCEP）と提案法（PHASOR）の音素毎の認識率の比較を示す。無声子音/t/, /ch/, /ts/ で認識率の劣化が見られるが、他の音素では PHASOR により認識率が向上していることがわかる。また、2 倍以上認識率が向上している音素としては、/b/, /g/, /m/, /n/, /N/, /w/, /r/, /p/, /h/ がある。

図 6 は PHASOR, CMN（ケプストラム平均正規化法）、SS（スペクトル減算法）を比較する。デルタケプストラムはいずれの場合も併用している。この図により PHASOR は従来の雑音除去方法である CMN や SS よりも高い認識率が得られていることがわかる。また、PHASOR と CMN を組み合わせることにより、さらに認識率を向上できる。

4.2 不特定話者音素認識

不特定話者音素認識実験は日本語音韻バランス 216 単語データベースについて行なった。話者は男性女性各 10 名ずつである。音素 HMM は性別のモデルとした。音素モデルは 10 名の同性のデータベースの

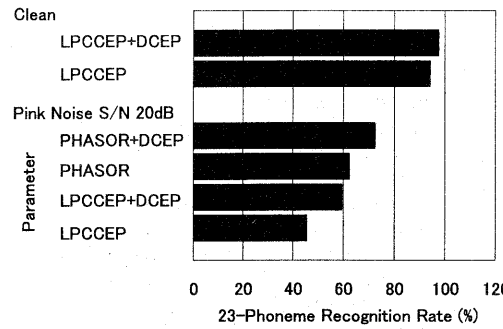

図 3 特定話者 23 音素認識実験結果（18 子音＋5 母音）

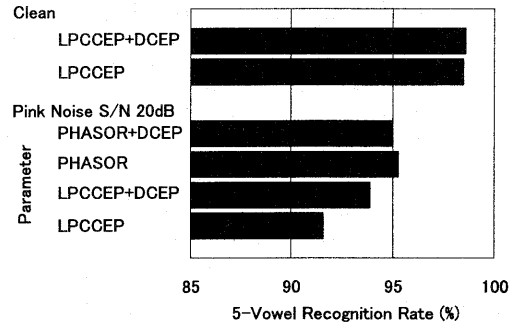

図 4 特定話者 5 母音認識実験結果

内、9 名の音声を用いて作成し、残りの 1 名の音声を認識した。テスト用話者を 10 名から順に選び、10 回の実験を行ない平均値を求めた。

図 7と図 8 は、それぞれ男性と女性の 23 音素認識実験結果である。図では特定話者音素認識実験と同様に、雑音が無い場合での結果を参考として示し、雑音付加音声の認識結果を従来法と PHASOR で比較している。付加雑音は特定話者認識実験の場合と同様、ピンク雑音で SNR は 20dB である。デルタケプストラムを用いない場合と用いる場合を組で示す。これらの図から、特定話者の場合と同様に、不特定話者においても提案法である PHASOR の認識率が高いことがわかる。

5. むすび

本報告では、雑音に対して頑健な新しいスペクトル推定法 PHASOR を提案した。本方法の特徴は有声部の繰り返し波形をフレーム内での位相同期加算により求めることが特徴である。PHASOR による雑音抑制効果を定量的に述べ、音素認識実験によりその効果を調べた。特定話者、不特定話者音素認識実験の結果、PHASOR により、大幅に雑音下音素認識率を向上できることを示した。また、PHASOR は、

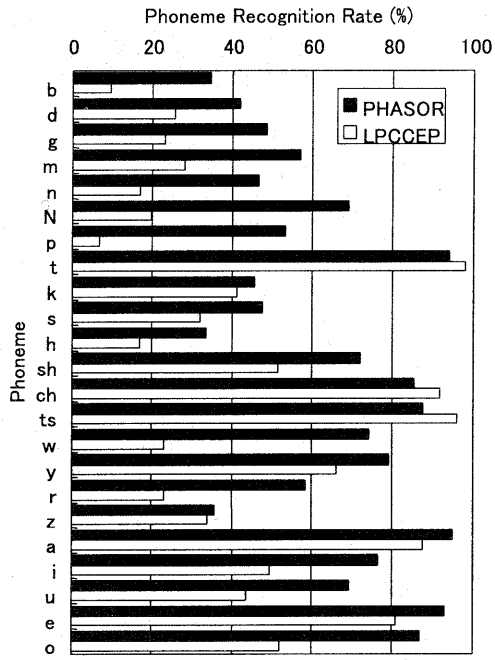

図 5　SNR=20dB でピンク雑音を付加した音声に対する PHASOR と従来法の音素ごとの認識率の比較。LPCCEP: 従来法

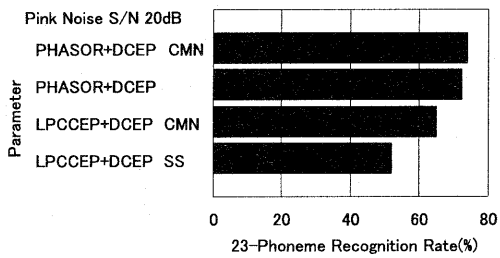

図 6　PHASOR, CMN（ケプストラム平均正規化法）, SS（スペクトル減算法）の雑音下 23 音素認識率の比較

従来よく用いられていたケプストラム平均正規化法 (CMN) や、スペクトル減算法 (SS) よりも高い雑音下音素認識率を示した。

文　献

[1] 石塚, 相川, “雑音下母音聴取における雑音のスペクトル構造の影響”, 信学技報, vol. SP2000-151, pp. 67–72 (2001-02).

[2] 石塚, 相川, “白色雑音・調波複合音下での母音知覚特性の比較”, 信学技報, vol. SP2001-44, pp. 23–28 (2001-07).

[3] 石塚, 相川, “雑音複合音下母音知覚における母音の振幅・基本周波数の時間変動の影響”, 音学講論, vol. I, pp. 447–448 (2001-10).

[4] K. Aikawa, “Speaker-independent speech recognition using micro segment spectrum estimation”, *ICSLP98*, vol. 4, pp. 1371–1374 (1998).

[5] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction”, *IEEE Trans. ASSP*, vol. 27, no. 2, pp. 113–120 (1979).

[6] F. Liu, A. Acero and R. Stern, “Efficient joint compensation of speech for the effects of additive noise and linear filtering”, *ICASSP92*, pp. 865–868 (1992).

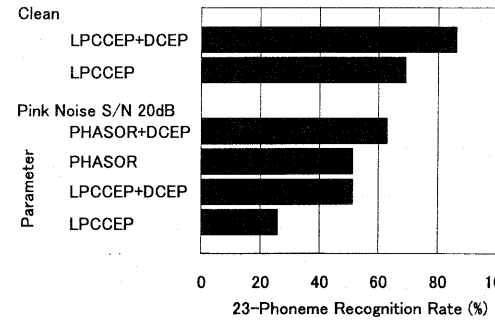

図 7　不特定話者 23 音素認識実験結果（男声）

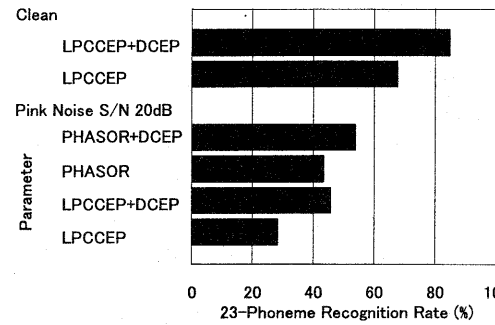

図 8　不特定話者 23 音素認識実験結果（女声）

謝　辞

熱心に議論していただいたマルチモーダル対話研究グループのメンバーに感謝する。