Internal Use Only(非公開)

TR-I-0316

雑音環境下文節認識実験による WLR・WGD・SGDS距離尺度の比較

Comparison among WLR,WGD and SGDS distance measure under noisy environment with HMM-LR

大倉計美　　杉山雅英
Kazumi OHKURA　Masahide SUGIYAMA

1993.3.3

概要

本報告では、雑音環境下における、特徴パラメータの堅牢性をWLR,WGD,SGDS距離尺度の比較およびケプストラムのユークリッド距離との比較により検討した結果を示す。SGDSはWLRとほぼ同等の性能を示すことと、ケプストラムは雑音に対して非常に弱いパラメータであることがわかる。WGDは距離の正定値性が証明されていない距離尺度であるが、WLRよりも雑音に対して堅牢である優れた距離尺度であることを確認した。

ATR自動翻訳電話研究所
ATR Interpreting Telephony Research Laboratories

© (株)ATR自動翻訳電話研究所　1993
© 1993 by ATR Interpreting Telephony Research Laboratories

目次

1 はじめに

雑音環境下の音声認識方法には、ノイズサブトラクション[1]の様な重畳雑音を除去する方法[2]-[5]や、雑音重畳空間を雑音の無い空間へのマッピングすることにより雑音を抑圧する手法および、雑音に対して堅牢な特徴パラメータを用いる事によって雑音に対処する方法、等が考えられる。本報告では、堅牢な特徴パラメータを用いる事によって雑音に対処する方法、および異空間の間のマッピング[7]を用いた耐雑音性の検討を行う。第2章において、雑音環境下における、特徴パラメータの堅牢性をWLR[9]、WGD[11]、SGDS距離尺度[12]、ケプストラムのユークリッド距離との比較により検討した結果を示す。第3章において、第2章の実験結果より耐雑音性の高かったWGDとATRで用いているWLRを距離尺度に選び、コードブックマッピング手法を用いた雑音抑圧手法について述べる。

2 距離尺度の概要

2.1では、現在ATRで用いているパラメータについて説明する。2.2では、WGD、2.3でSGDSについて説明する。2.4では、実験概要と実験結果を示す。

2.1 ATRで用いている距離尺度

現在、離散分布型HMMによる認識では、特徴量として、相関係数、Δケプストラム及びパワーを用いている。VQコード列を算出する場合の距離計算式として、以下に示す3種類の距離を用いている。[9][10]

$$d_{WLR} = \sum_{j=1}^{N}(r_j - r'_j)(c_j - c'_j) \tag{1}$$

$$d_{\Delta CEP} = \sum_{j=1}^{N}(\Delta c_j - \Delta c'_j)^2 \tag{2}$$

$$\Delta c_j(k) = [\sum_{n=-8}^{8}(c_{j+n}(k)n)]/\sum_{n=-8}^{8} n^2$$

$$d_{POW} = p/p' + p'/p - 2 \tag{3}$$

ただし、r_j、c_j、Δc_j、pはそれぞれfの第j次のＬＰＣ自己相関係数、第j次のLPCケプストラム係数、17フレーム　（５１ｍｓ）のc_jの回帰係数、パワーを表す。r'_j、c'_j、$\Delta c'_j$、p'は、gを表す。NはWLR尺度およびΔケプストラムの計算打切り次数を表す。

2.2 WGD距離尺度の概要

信州大学の松本等が提案している距離尺度であり、式(4)で表される。WGDは高次のパラメータのウエイトが大きくなっており、スペクトルの傾きを抑え、ピークを強調するかたちになっている。特長として、スペクトルの全体的傾斜の変動に強く、低次ホルマントに大きな感度を持つことが示されている[11]。

$$d_{WGD} = \sum_{j=1}^{N} j(r_j - r'_j)(c_j - c'_j) \tag{4}$$

式(4)中の記号の意味は、式(1)に示したものと同様である。

2.3 SGDS距離尺度の概要

SGDS距離尺度は次式で表される距離尺度である[12]。

$$d_{SGDS} = \sum_{j=1}^{N} (w_j(c_j - c'_j))^2 \tag{5}$$

式(5)中の記号の意味は、式(1)に示したものと同様である。また、w_j は次式で表される平滑化重み関数である。

$$w_j = j^s exp(-j^2/(2\tau^2)) \tag{6}$$

ここで、j^s の項は c_j に対する重みの度合いで、s が大きくなるに従って重みを増加し、低次のケプストラム成分の占める割合を少なくする。そのため、スペクトルの先鋭化と、スペクトル概形成分を取り除く作用がある。次に $exp(-j^2/(2\tau^2))$ の項は j の大きいところでの重みを抑え、周波数領域における平滑化を行う作用が有り、τ が小さいほど、周波数分解能が低下する。本実験では $s = 1.0$、$\tau = 12.0$ とした。

2.4 実験概要と結果

2.4.1 実験概要

評価はHMM-LR[13]を用いた文節認識実験で行なう。話者は男性(MAU)である。HMMの学習には雑音を重畳していない5240単語から切り出した音素を用いた。入力文節数は279文節。音響分析は、12kHzサンプリング、分析窓長21.3ms、フレーム周期3ms、14次のLPC分析を行い、コードブックとしてWLR（WGD）距離尺度より求めた自己相関係数（256）、LPCケプストラムの回帰係数である Δケプストラム(256)、正規化パワー（64）を用いた。ただし、ケプストラムとSGDSの場合はそれぞれを256サイズのコードブックとして持つ。雑音重畳資料はアナログ雑音発声器で作成した有色雑音を計算機上で音声にSNRが30dB, 20dBになるように重畳したものを用いた。各距離尺度(WLR, SGDS, WGD)の打切り次数は16である。

2.4.2 実験結果

実験結果を表1に示す。表1より、SGDSはWLRよりも優れた耐雑音性を示すことと、ケプストラムは雑音に対して非常に弱いパラメータであることがわかる。WGDは距離の正定値性が証明されていない距離尺度であるが、4種類のパラメータのうち最も雑音に対して堅牢である優れた距離尺度であることがわかる。また、SGDは雑音の無い環境ではケプストラムと同程度の認識率であるが、雑音環境下では、WLRよりも高い認識性能を示すことが分かる。

3　コードブックマッピングによる雑音抑圧

本章では、第2章の実験結果より耐雑音性の高かったWGDとATRで用いているWLRを距離尺度に選び、コードブックマッピング手法[6]を用いた雑音抑圧手法について述べる。

3.1　ヒストグラムに基づく話者適応方式

本方法は、ベクトル量子化が特徴空間の離散表現になっていることを利用し、有限個の離散点の話者間の関係を見いだすことにより、話者適応を行うものである。

話者間の対応関係は、未知話者のコードブックを用いてベクトル量子化された未知話者の学習音声と、標準話者のコードブックを用いてベクトル量子化された標準話者の学習音声間で、非線形マッチング(DTW)を行うことにより求められ、対応付けヒストグラムにより表現される。以下に、本方法のアルゴリズムを示す。

(話者適応アルゴリズム)

[学習]

ステップ1：　未知話者（A）の学習単語より未知話者のベクトル量子化コードブック $\{v_{iA}\}$ を作成する。

ステップ2：　未知話者の学習音声を $\{v_{iA}\}$ を用いてベクトル量子化する。

ステップ3：　標準話者（B）のコードブック $\{v_{jB}\}$ でベクトル量子化した標準話者の学習単語と未知話者の学習単語間でDTWを行い、最適パスを求める。

ステップ4：　DTWの最適パスに従い、$\{v_{iA}\}$ と $\{v_{jB}\}$ 間のベクトルの対応回数を求め、対応付けヒストグラム（h_{ij}）を求める。

ステップ5：　対応付けヒストグラムの値を重みとし、未知話者の空間を標準話者の空間に写像するための変換コードブック $\{v_{iA-B}\}$ を求める。

ステップ6：　$\{v_{iA}\}$ を $\{v_{iA-B}\}$ に入れ換える。

ステップ7：　DTW時の距離が収束していないならば、3へ戻る。

(ヒストグラムに基づく出力確率の変換)

認識時には学習時に求めたヒストグラム h_{ijn} を両者のコードベクトルにおける対応付けの確からしさと見なし、次式に示すように標準話者のHMMのコードベクトルの出力確率 b_{kjn} と各特徴パラメータ毎に求まったヒストグラム h_{ijn} の積をとることにより、標準話者のコードベクトルの出力確率 $\omega_t(k)$ を変換することにより認識を行う。

$$\omega_t(k) = \prod_{n=1}^{P} \left(\sum_{j=1}^{M_n} \left(\sum_{i=1}^{C_n} u_{in} h_{ijn} \right) b_{kjn} \right)$$

ただし、nは特徴パラメータ番号(CEP,ΔCEP,POW)、tは時刻、kはHMMのステート番号、iは入力コードベクトルの番号を表す。

3.2 実験概要と結果

3.2.1 実験概要

評価は HMM-LR[13] を用いた文節認識実験で行なう。雑音適応のみを行なった場合と、話者適応と雑音適応を同時に行なった場合の 2 実験により評価を行なう。適応には、音素バランス 216 単語の先頭から 100 単語を用いた。入力話者は男性 (MAU) 1 名である。雑音適応のみを行う場合は、MAU の 1 回目発声の音声と雑音を重畳した 2 回目発声音声間でマッピングを行なう。話者適応を行なう場合の標準話者は男性 (MHT) である。この場合は、雑音を重畳した 1 回目発声の音声と MHT の音声間でマッピングを行なう。また、コードブックマッピングの環境変化に対する堅牢性を調べるためにヒストグラムをもとめた SNR とはことなる SNR の音声に対する文節認識実験も行なう。HMM の学習には雑音を重畳していない 5240 単語から切り出した音素 (各音素最大 400 サンプル) を用いた。入力文節数は 279 文節。音響分析は、12kHz サンプリング、分析窓長 21.3ms、フレーム周期 3ms、14 次の LPC 分析を行い、コードブックとして WLR および WGD 距離尺度より求めた自己相関係数 (256) 、LPC ケプストラムの回帰係数である Δ ケプストラム (256)、正規化パワー (64) を用いた。コードブックは、音素バランス 216 単語を全て使用して作成した。雑音重畳資料はアナログ雑音発声器で作成した有色雑音を計算機上で音声に SNR が 30dB, 20dB になるように重畳したものを用いた。WLR, Δcep の計算打切り次数は 16 である。距離尺度の詳細は 2.1 および 2.2 で述べたものと同様である。

コードブックマッピングは、DTW によるコードベクトルの対応づけにより実現される。この対応づけを行なう場合に、重みを掛けている WGD 距離尺度 (d_{WGD}) は、WLR 距離尺度に基づく距離 (d_{WLR}) よりも大きな距離値を示す。このため DTW の距離計算時は次式の様に補正を行なった。ただし、本補正値は最適なものではなく、今後検討する必要がある。

$$Dist = 0.25 d_{WGD} + 0.3 d_{\Delta CEP} + 0.01 d_{POW}$$

$$Dist = d_{WLR} + 0.3 d_{\Delta CEP} + 0.01 d_{POW}$$

3.3 実験結果

1. 雑音抑圧性能の評価

コードブックマッピングを雑音抑圧に用いた場合の結果を表 2、表 4に示す。また、HMM-LR システムでは、音声の切り出しを無音区間の HMM を用いて行なっており、雑音環境下における音声区間の切り出し精度の向上を期待し、無音区間 HMM を雑音環境下で学習した場合の認識率も合わせて示す。

表より、無音区間 HMM の雑音環境下学習は、簡単でしかも有効な手段であることが分かる。また、雑音環境下において、コードブックマッピング手法の認識率は、雑音処理を行なわない場合の認識率を上回っており、コードブックマッピング手法の雑音抑圧への応用の有効性がわかる。距離尺度に WLR を用いた場合と WGD を用いた場合とでは、認識率に違いは殆んど見られなかった。また、SNR による認識率の変化も同じ傾向を示すことが分かるので、以下では、WLR 距離尺度の認識結果について検討する。無音区間 HMM の雑音環境下学習を行

なったものと、コードブックマッピング手法を比較すると、SNR = 20 dB においてそれぞれ 49.1%、55.6% という認識率である。この結果より、SNR = 20 dB では、認識率の低下の原因として音声の切り出し誤りの他に音声特徴そのものが雑音により崩れているが、マッピングによる手法はこの崩れを学習していると考えられる。

更に、SNR = 30 dB で求めたヒストグラムを用いて、SNR = 20 dB の音声を認識した場合、61.3% の文節認識率が得られている。この認識率は、SNR = 20 dB で求めたヒストグラムを用いて SNR = 20 dB の音声を認識した場合の認識率 (55.6%) よりも高い文節認識率である。この原因の 1 つとして、SNR = 20 dB においては、DTW のパスが不正確でヒストグラムが正確に求められていないことが考えられる。しかし、SNR = 30 dB 程度の SNR でヒストグラムを求めれば、SNR = 20 dB の音声をカバーできることが明らかになった。

2. 話者適応と雑音抑圧性能の評価

コードブックマッピングを話者適応と雑音抑圧に用いた場合の結果を表 3、表 5 に示す。表より、雑音環境下では話者適応と雑音適応を同時に行なった認識率が特定話者の認識率を上回っており、話者適応と雑音抑圧を同時に行なえることが分かる。また、クロス SNR マッピングでは、"雑音抑圧性能の評価" と同様に、SNR = 30 dB で求めたヒストグラムを用いて、SNR = 20 dB の音声を認識した場合に高い認識率が得られている。

以上の 2 実験より、コードブックマッピング手法を雑音と話者性を同時に扱う、環境適応へ応用できることが明らかになった。また、WLR と WGD 距離尺度の違いによるマッピング性能は、ほぼ同等の性能であることが分かった。ただし、DTW の距離計算時に行なった補正値の検討により認識率が向上する可能性があり、今後検討する必要がある。

4 あとがき

本報告では、HMM-LR を用いた雑音環境下での文節認識実験を通して、WLR, WGD, SGDS, CEP のユークリッド距離の 4 種類の距離尺度の耐雑音性能を検討した。結果としてホルマントを強調した距離尺度である、WLR, WGD, SGDS がケプストラムよりもかなり優れた距離尺度であることを確認できた。また、話者適応手法を異空間のマッピングを実現する手法として捉え、環境と話者性を同一に扱うことにより環境適応への応用を検討した。結果として雑音環境下では話者適応と雑音適応を同時に行なった認識率が特定話者の認識率を上回り、話者適応と雑音抑圧を同時に行なえることが明らかになった。

参考文献

[1] Steven F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. on ASSP, Vol.ASSP-27, No.2, pp.113-120 (Apr. 1979).

[2] Jae S. Lim, Alan V. Oppenheim and Louis D. Braida, "Evaluation of an Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition," IEEE Trans. on ASSP, Vol.ASSP-26, No.4, pp.354-358 (Aug. 1978).

[3] B. Widrow, J. R. Glover and et al., "Adaptive Noise Cancelling: Principle and Applications," Proc. of IEEE, Vol.63, No.12, pp.1692-1716 (Oct. 1975).

[4] M. R. Sumber, "Adaptive Noise Canceling for Speech Signals," IEEE Trans. ASSP, Vol. ASSP-26, No.5, pp.419-423 (Aug. 1979).

[5] A. P. Varga and R. K. Moore, "Hidden Markov Model Decomposition of Speech and Noise," Proc. of ICASSP90, pp.845-848 (Apr. 1990).

[6] S. Nakamura and K. Shikano, "Speaker Adaptation Applied to HMM and Neural Networks," proc. of ICASSP89, pp.89-92 (May 1989).

[7] K. Ohkura and M. Sugiyama, "Speech Recognition in a Noisy Environment Using a Noise Reduction Neural Network and a Codebook Mapping Technique," Proc. of ICASSP91, pp.929-932 (May 1990).

[8] T. Hanazawa, T. Kawabata and K. Shikano, "Recognition of Japanese Voiced Stops Using Hidden Markov Models," Journal of ASJ, Vol.10, No.10, pp.776-785 (Oct.1989).

[9] M. Sugiyama and K. Shikano, "LPC Peak Weighted Spectral Matching Measures," Trans. of IECE, Vol.J64-A, No.5, pp.409-416 (May 1981) (in Japanese).

[10] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," IEEE Trans. on ASSP, Vol.ASSP-34, pp.52-59 (Feb. 1986).

[11] H. Matsumoto and H. Imai, "Comparative Study of Various Spectrum Matching Measures on Noise Robustness," Proc. of ICASSP86, pp.769-772 (Apr. 1986).

[12] F. Itakura and T. Umezaki, "Distance Measure for Speech Recognition Based on The Smoothed Group Delay Spectrum," Proc. of ICASSP87, pp.1257-1260 (Apr. 1987).

[13] K. Kita, T. Kawabata and H. Saito, "HMM Continuous Speech Recognition Using Predictive LR Parsing," Proc. of ICASSP89, pp.703-706 (May 1989).

表 1: 各距離尺度による文節認識率

SNR	WLR	WGD	SGDS	CEP
∞ dB				
1 位	85.3%	85.3%	84.6%	84.9%
~2 位	96.4%	96.1%	93.9%	95.3%
~3 位	97.5%	97.5%	96.8%	97.1%
~4 位	98.2%	98.6%	97.8%	98.2%
~5 位	99.6%	99.3%	98.2%	98.2%
30 dB				
1 位	70.6%	75.3%	74.2%	65.6%
~2 位	86.0%	92.5%	87.1%	76.3%
~3 位	88.9%	95.7%	93.2%	81.7%
~4 位	92.8%	97.5%	95.0%	84.2%
~5 位	94.3%	98.2%	96.1%	86.0%
20 dB				
1 位	36.9%	49.5%	45.2%	29.0%
~2 位	48.5%	62.4%	55.2%	35.8%
~3 位	55.2%	69.5%	59.1%	41.2%
~4 位	58.4%	72.0%	61.3%	47.3%
~5 位	60.6%	74.2%	64.5%	47.3%

表 2: コードブックマッピング (mau→mau)
距離尺度:WLR,Δcep,POW

		適応を行なったSNR ∞dB	適応を行なったSNR 30dB	適応を行なったSNR 20 dB	適応は行なわず無音区間HMMを雑音環境下で学習	雑音処理を行なわない場合(MAU)
入	SNR=∞dB					
	1位	*81.7%*	72.4%	62.0%	85.3%	85.3%
	~2位	*90.7%*	86.7%	74.9%	96.4%	96.4%
	~3位	*94.6%*	91.8%	83.5%	97.5%	97.5%
	~4位	*96.1%*	93.9%	87.1%	98.2%	98.2%
	~5位	*97.8%*	95.7%	88.9%	99.6%	99.6%
力	SNR=30 dB					
	1位	72.8%	*72.8%*	61.3%	77.8%	70.6%
	~2位	86.7%	*87.5%*	74.2%	90.3%	86.0%
	~3位	92.1%	*91.4%*	81.4%	92.8%	88.9%
	~4位	94.6%	*92.8%*	86.7%	95.0%	92.8%
	~5位	95.3%	*94.3%*	88.5%	95.7%	94.3%
	SNR=20 dB					
	~1位	38.0%	61.3%	*55.6%*	49.1%	36.9%
	~2位	48.4%	75.3%	*71.0%*	58.8%	48.5%
	~3位	56.3%	79.9%	*76.7%*	63.1%	55.2%
	~4位	57.3%	82.4%	*78.5%*	68.1%	58.4%
	~5位	59.9%	84.9%	*81.4%*	69.2%	60.6%

表中のイタリック文字は、
雑音適応時と認識時のSNRが同じ場合の認識率を示す

表 3: コードブックマッピング (mau→mht)
距離尺度:WLR,Δcep,POW

		適応を行なった SNR ∞dB	適応を行なった SNR 30dB	適応を行なった SNR 20 dB	特定話者認識率
入力	SNR=∞dB				
	1 位	*79.9%*	76.0%	63.8%	85.3%
	~2 位	*92.5%*	88.2%	78.9%	96.4%
	~3 位	*96.4%*	91.8%	84.9%	97.5%
	~4 位	*96.8%*	93.2%	90.0%	98.2%
	~5 位	*97.8%*	93.9%	91.0%	99.6%
	SNR=30 dB				
	1 位	74.6%	*71.7%*	67.0%	70.6%
	~2 位	88.5%	*87.1%*	81.7%	86.0%
	~3 位	93.2%	*91.0%*	85.7%	88.9%
	~4 位	95.3%	*92.8%*	88.5%	92.8%
	~5 位	96.4%	*94.3%*	91.0%	94.3%
	SNR=20 dB				
	1 位	36.6%	67.4%	*58.8%*	36.9%
	~2 位	50.2%	82.4%	*71.3%*	48.5%
	~3 位	57.0%	87.8%	*75.3%*	55.2%
	~4 位	58.8%	89.2%	*76.7%*	58.4%
	~5 位	61.6%	90.3%	*79.2%*	60.6%

表中のイタリック文字は、
話者適応時と認識時の SNR が同じ場合の認識率を示す

表 4: コードブックマッピング (mau→mau)
距離尺度:WGD,Δcep,POW

		適応を行なったSNR ∞dB	適応を行なったSNR 30dB	適応を行なったSNR 20 dB	適応は行なわず無音区間HMMを雑音環境下で学習	雑音処理を行なわない場合
入	SNR=∞dB					
	1位	*82.1%*	78.5%	58.1%	85.3%	85.3%
	~2位	*92.5%*	88.2%	77.4%	96.1%	96.1%
	~3位	*95.7%*	93.2%	82.1%	97.5%	97.5%
	~4位	*96.1%*	93.9%	85.3%	98.6%	98.6%
	~5位	*97.1%*	95.0%	87.5%	99.3%	99.3%
力	SNR=30 dB					
	1位	76.7%	*77.4%*	62.7%	78.5%	75.3%
	~2位	88.5%	*85.7%*	76.0%	92.5%	92.5%
	~3位	92.8%	*90.3%*	81.7%	94.3%	95.7%
	~4位	94.6%	*91.4%*	85.3%	97.1%	97.5%
	~5位	96.1%	*92.8%*	87.5%	97.1%	98.2%
	SNR=20 dB					
	~1位	53.8%	63.4%	*53.4%*	52.0%	49.5%
	~2位	62.0%	72.8%	*65.2%*	62.0%	62.4%
	~3位	67.7%	76.7%	*70.6%*	67.4%	69.5%
	~4位	71.0%	80.6%	*74.2%*	69.2%	72.0%
	~5位	75.3%	82.4%	*76.3%*	71.3%	74.2%

表中のイタリック文字は、
雑音適応時と認識時のSNRが同じ場合の認識率を示す

表 5: コードブックマッピング (mau→mht)
距離尺度:WGD,Δcep,POW

		適応を行なった SNR ∞dB	適応を行なった SNR 30dB	適応を行なった SNR 20 dB	特定話者認識率 (MAU)
入力	SNR=∞dB				
	1 位	*76.7%*	74.9%	61.6%	85.3%
	~2 位	*91.8%*	87.5%	77.1%	96.1%
	~3 位	*95.0%*	91.0%	83.9%	97.5%
	~4 位	*96.4%*	93.5%	86.7%	98.6%
	~5 位	*96.8%*	93.9%	88.2%	99.3%
	SNR=30 dB				
	1 位	74.9%	*72.0%*	66.7%	75.3%
	~2 位	89.6%	*86.0%*	78.9%	92.5%
	~3 位	93.9%	*90.3%*	84.6%	95.7%
	~4 位	95.0%	*91.8%*	86.0%	97.5%
	~5 位	95.7%	*92.5%*	87.8%	98.2%
	SNR=20 dB				
	1 位	45.5%	68.1%	*56.3%*	49.5%
	~2 位	60.2%	79.2%	*68.1%*	62.4%
	~3 位	65.6%	83.2%	*72.8%*	69.5%
	~4 位	68.8%	85.7%	*73.8%*	72.0%
	~5 位	69.9%	86.7%	*75.3%*	74.2%

表中のイタリック文字は、
話者適応時と認識時の SNR が同じ場合の認識率を示す