論文 / 著書情報
Article / Book Information

論題(和文)	CSM強度対を用いた音声認識
Title(English)	
著者(和文)	五十川賢造, 篠田浩一, 嵯峨山茂樹
Authors(English)	Koichi Shinoda
出典(和文)	日本音響学会平成14年春季研究発表会講演論文集, Vol. 1-5-4, No. , pp. 7-8
Citation(English)	, Vol. 1-5-4, No. , pp. 7-8
発行日 / Pub. date	2002,

1—5—4

CSM強度対を用いた音声認識*

◎五十川賢造　篠田浩一　嵯峨山茂樹（東大、情報理工）

1　はじめに

本稿では、音声認識の特徴量として、線スペクトル対 (LSP)[1] と複合正弦波モデル (CSM) の利用を検討する。

LSP 周波数は補間・量子化特性に優れ、携帯電話などの音声圧縮に広く用いられていることから、音声認識にも有用な特徴量である期待が持たれ、ケプストラムと同等以上の性能が確認されている [2, 3, 4, 5]。これらの研究では、LSP 周波数自体、あるいはその時間微分を音声認識の特徴量として用いていた。これに対し本稿では、LSP 周波数から強度に相当する量を導くことを考える。

特徴ベクトルとして固定された周波数帯域に対するフィルタ出力を、複数個扱う場合を考える。この場合、十分な解像度の周波数領域の情報を得るにはフィルタ数を多く取らねばならないという問題がある。しかし、入力されたスペクトルに応じて強度が小さくスペクトル包絡が平坦な周波数帯域では解像度を下げ、強度が大きくスペクトル包絡が平坦でない周波数帯域では解像度を上げる様に自動的に適応するフィルタバンクが構成できるならば、より効率のよい情報の取り扱いが可能であることが期待される。また、各フィルタが対応する周波数帯域が入力スペクトルに対して最適に定められるならば、フォルマントの移動等による話者性の影響ををを少なくできる可能性がある。

本稿では、LSP 周波数から得られる強度に相当する量であり、上に述べたフィルタバンクの発想に適合する特徴量として、LSP 周波数間隔と CSM 強度対の 2 つの特徴量を紹介する。LSP 周波数間隔は LSP 周波数によるパワー周波数応答関数のピークと密接な関係があり、CSM 強度対は CSM を介して LSP 周波数と相補的な関係にある。さらに実験により、2 つの特徴量の実際の音声認識への利用の可能性を探る

2　LSP 周波数間隔

分析フレームごとに音声信号を分析して得られる p 次 LSP 周波数を $\{\omega_1, \omega_2, \cdots, \omega_p\}$[rad] とすると、隣接する LSP 周波数が近接する区間ではスペクトルが大きな値を取ることは、経験的に知られている。もう少し詳しく見てみよう。LSP 周波数により決定される全極型フィルタのパワー周波数応答関数 $|G(e^{j\omega})|^2$ は以下の式で表される [1]。

$$|G(e^{j\omega})|^2 = \frac{1}{2^p\left[\sin^2\frac{\omega}{2}\prod_{i=2,4,\cdots}(\cos\omega-\cos\omega_i)^2+\cos^2\frac{\omega}{2}\prod_{i=1,3,\cdots}(\cos\omega-\cos\omega_i)^2\right]}$$

LSP 周波数が近接する区間 (ω_k, ω_{k+1}) では、A, B, A', B', A'', B'' を正の定数として、

$$|G(e^{j\omega})|^2 \approx \frac{1}{A(\cos\omega-\cos\omega_k)^2+B(\cos\omega-\cos\omega_{k+1})^2} \approx \frac{1}{A'\sin^2\frac{\omega-\omega_k}{2}+B'\sin^2\frac{\omega-\omega_{k+1}}{2}}$$

* "Speech Recognition using CSM Intensity Pair" by Kenzo Isogawa, Koichi Shinoda and Shigeki Sagayama, Department of Information Physics and Computing, Graduate School of Information Science and Technology, The University of Tokyo.

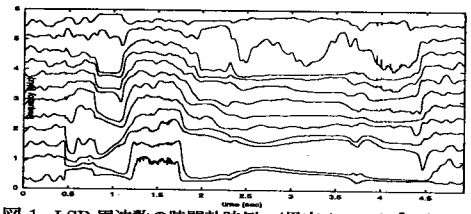

図 1: LSP 周波数の時間軌跡例。(男声 (MAU)「こしらえる」。各軌跡は下から順に $\omega_1, \omega_2, \cdots$ に対応する。)

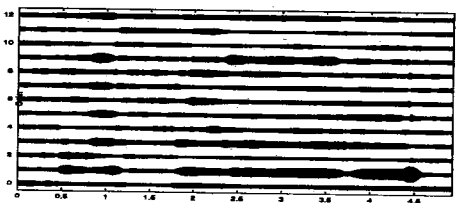

図 2: LSP 周波数間隔の逆数の時間パターン例。(男声 (MAU)「こしらえる」。下から順に $\delta_0, \delta_1, \cdots$ に対応し、線幅は逆数に比例する。)

$$\approx \frac{1}{A''(\omega-\omega_k)^2+B''(\omega-\omega_{k+1})^2}$$

と近似できる。このことから、LSP 周波数間隔 $\delta_k = \omega_{k+1}-\omega_k$ は近似的にスペクトルピークの半値幅とみなすことができる。共振系における半値幅とスペクトルピークの関係から、スペクトルのピーク値はおおむね $1/\delta_k^2$ に比例すると考えて良い。

本稿では、$\omega_0 = 0, \omega_{p+1} = \pi$ として、$k = 0, \cdots, p$ について $\delta_k = \omega_{k+1}-\omega_k$ を LSP 間隔 (δLSP) と呼び、その対数値を音声認識の特徴量として用いる。LSP 周波数軌跡を図 1 に、δLSP の逆数の時間パターンの例を図 2 に示す。

LSP 周波数が、パワースペクトルに対して、情報の多い箇所に適応的に配置され (CSM の理論に基づいて説明できる)、$-2\log\delta$LSP は対数スペクトルに対応することから、直接に周波数を情報として用いる場合に比べ、(1) フォルマントの移動に対して頑健、(2) 情報量の少ない (強度の低い) 周波数付近の情報を圧縮できる、という性質を持つことが期待される。同時に、これらは性能劣化の要因ともなり得るので、次々章にて実験検証する。

3　CSM 強度対 (LSP 強度)

LSP と密接な関係にある CSM[6, 7] は、音声信号を複数個の正弦波の和でモデル化するものである。モデルのパラメタは各正弦波の周波数 ω_i と強度 m_i である。モデルの正弦波の周波数は 4 種の「拘束条件」のもとで求めることが可能であり、全て異なった解となる。4 種の条件とは、モデルの全正弦波の周波数に拘束がない場合、モデルが直流を含む場合、モデルがナイキスト周波数 (π) の正弦波を含む場合、モデルが直流とナイキスト周波数の正弦波の両方を持つ場合である。分析次数 p が偶数・奇数の場合それぞれに 2 種の拘束条件のもとでモデルのパラメタを求める。

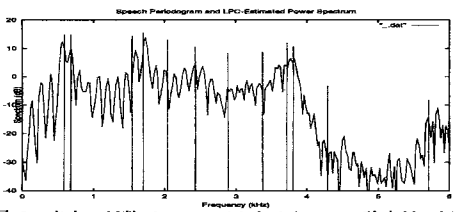

図 3: 音声の対数パワースペクトルと CSM 強度対の例。(縦線の位置は CSM 周波数対、高さは CSM 強度対 (対数) を表わす)

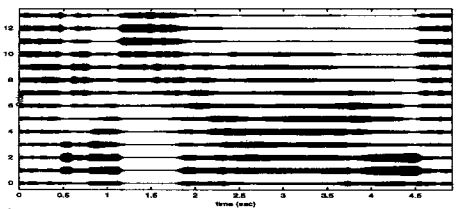

図 4: 12 次 CSM 強度対の時間パターンの例。(男声 (MAU) 「こしらえる」。下から順に CSM 強度対 $m_0, m_1, \cdots$ に対応し、線幅は $\sqrt{m_i}$ を表す。)

たとえば $p=$ 偶数のとき、観測された信号の自己相関関数 v_τ を用いて、$\tau = 0, 1, \cdots, p$ について CSM 自己相関方程式を解く。

$$m_0 + \sum_{i=2,4,\cdots,p} m_i \cos\tau\omega_i = \sum_{i=1,3,\cdots,p-1} m_i \cos\tau\omega_i + (-1)^\tau m_{p+1} = v_\tau$$

この式には必ず解が存在し、CSM 周波数対 $\{\omega_0 = 0, \omega_1, \cdots, \omega_p, \omega_{p+1} = \pi\}$ および CSM 強度対 $\{m_0, m_1, \cdots, m_p, m_{p+1}\}$ が得られる。CSM 周波数対は LSP 周波数と同一であり、またすべて $m_i > 0$ であることが証明されている [7, 8]。LSP の周波数領域での表現が LSP 周波数ならば、CSM 強度対は LSP 強度とでも呼ぶべき量となる。

本稿での意図に沿って音声認識の特徴量を検討する。CSM 強度対は LSP 周波数と同等の情報を担うが、その表現領域は周波数でなく強度である。図 3 に音声のパワースペクトルとそれに対応する CSM 周波数対 (LSP 周波数) および強度対の例を示す。CSM 強度対は、Gauss-Jacobi 求積法の意味 (あるいは LSP の意味) で最適な帯域区分を行ったスペクトルの強度 [7] という意味もあり、適応的なフィルタバンク出力のようなものと理解できる。CSM 強度対の時間軌跡の例を図 4 に示す。少数の線でスペクトログラム様の情報を表現できることがわかる。CSM 強度対の対数値 $\log m_i$ を音声認識の特徴量として用いる。

4 音声認識による評価

単一ガウス分布音素 HMM による単語認識実験を行った。実験条件を表 1 に示す。表 2 の 4 つの特徴量に対して特定話者の認識を行った結果、表 3 の認識率を得た。認識率では、δLSP、CSM 強度対ともに、MFCC には及ばなかったが、LSP 周波数をそのまま用いた場合より高い性能を得た。

次に複数話者モデルから、特定話者モデルに移行した際の認識誤り減少率を表 4 に示す。複数話者モデルとして男声モデル (男性 5 人のデータで学習)、女声モデル (女性 4 人のデータで学習) を使用した。男声モデルのみではあるが、δLSP、CSM 強度対ともに MFCC より低い認識誤り減少率を得た。この結果は複数話者モデルと特定話者モデルの性能が近いことを示しており、話者に関する特徴量の正規化のような効果が得られていると解釈できる。

表 1: 音声認識実験条件 (各実験で共通)

単語データ	ATR データベース A セット 話者 : mau,mht,mmy,msh,mtk ffs,fkn,fms,fyn
標本化周波数	20kHz
特徴量次元	12 次
分析窓	窓幅 30msec、シフト 10msec
音響モデル	状態数 3、混合数 1
共分散行列	対角共分散
学習データ	奇数番目 2620 単語
辞書登録単語数	偶数番目 2620 単語
triphone 状態数	1700-1800

表 2: 特定話者認識実験の特徴量 (LSP 分析次数, p)

LSP 周波数	$\omega_1, \cdots, \omega_{12}$; $p = 12$
δLSP	$\log\delta_0, \cdots, \log\delta_{11}$; $p = 11$
CSM 強度対	$\log m_0, \cdots, \log m_{11}$; $p = 10$
MFCC	12 次元 (比較用)

表 3: 特定話者単語認識率 [%](男性 2 名, 女性 2 名の平均)

model	LSP 周波数	δLSP	CSM 強度対	MFCC
monophone	62.5	75.8	65.9	84.9
triphone	80.8	89.6	83.6	94.7

表 4: 認識誤り削減率 [%] (monophone)

モデル	δLSP	CSM 強度対	MFCC
男声 → 特定話者	31.2	30.3	40.8
女声 → 特定話者	37.2	38.3	36.6

5 終りに

LSP 周波数から導かれる 2 つの新しい音声特徴量、δLSP、CSM 強度対を提案した。認識性能と、話者依存性を調べるため、HMM による単語認識性能と、複数話者モデルから特定話者モデルに移行した際の誤り減少率を評価した。結果、単語認識性能は LSP 周波数を上回り、話者に関する特徴量の正規化の効果も確認できた。今後の課題としては、KL 展開などによりパラメータ間の相関を除去する、話者数を増やし話者性についてさらに考察する、ノイズに対する影響を調べる、認識率の向上を目指す、等を考慮中である。

参考文献

[1] 管村昇, 板倉文忠, "線スペクトル対 (LSP) 音声分析合成方式による音声情報圧縮," 電子通信学会論文誌, Vol. J64-A, No. 8, pp. 599–606, 1981.

[2] K. K. Paliwal, "A Study of Line Spectrum Pair Frequencies for Speech Recognition," *Proc. ICASSP'88*, Vol. 1, pp. 485–488, 1988.

[3] K. K. Paliwal, "A Study of LSF Representation for Speaker-Dependent and Speaker-Independent HMM-based Speech Recognition Systems," *Proc. ICASSP90*, Vol. 2, pp. 801–804, 1990.

[4] Fikret S. Gürgen, Shigeki Sagayama, and Sadaoki Furui, "Line Spectrum Pair Frequency-Based Distance Measures for Speech Recognition," *Proc. ICSLP90*, pp. 521-524, 1990.

[5] Seung Ho Choi, Hong Kook Kim, Huang Soo Lee, "LSP Weighting Functions Based on Spectral Sensitivity and Mel-Frequency Warping for Speech Recognition in Digital Communication," *Proc. ICASSP99*, Vol. 1, pp. 401–404, 1999.

[6] 嵯峨山茂樹, "CSM パラメータを用いたスペクトルマッチング尺度," 日本音響学会昭和 56 年春季研究発表会講演論文集, Vol. 2, pp. 513–514, 1976.

[7] 嵯峨山茂樹, 板倉文忠, "複合正弦波モデルによる音声スペクトルの解析," 電子通信学会論文誌, Vol. J64-A, No. 2, pp. 105–112, 1981.

[8] 嵯峨山茂樹, 板倉文忠, "線形予測符号化と複合正弦波モデルの対称性," 電子通信学会論文誌, Vol. J83-A, No. 11, pp. 1244–1255, 2000.