音声言語情報処理 34－1
(2000. 12. 21)

〔サーベイ〕

頑健な音声処理手法

-多元信号の統合に基づく音声処理-

武田 一哉

名古屋大学大学院 工学研究科

〒464-8603 名古屋市千種区不老町1
takeda@nuee.nagoya-u.ac.jp

あらまし 音声信号の頑健な分析を実現する方法として，同一音声をさまざまな方法で観測し，得られた信号を統合することでより安定した情報を抽出する方法が関心を集めている．部分帯域毎にモデル化と確率計算を行ない、その結果を用いる方法が最も多く検討されている方法である。手法のアイディアは斬新であるが、大きな基本性能の改善が得られるには至っていない．本稿では、このマルチストリーム型の音声処理手法に関して研究動向と従来手法との関連を論じる，

キーワード 音声認識, マルチストリーム

Robust Speech Recognition through Multiple Observations and their Integration

Kazuya TAKEDA

Graduate School of Engineering, Nagoya University

Furo-cho 1, Chikusa-ku, Nagoya 464-8603 JAPAN
takeda@nuee.nagoya-u.ac.jp

Abstract A new framework of robust speech processing, multiple observation and their integration is discussed. The basic scheme of the method is 1) capturing the properties of the speech through multi-stream signal representation, and 2) extracting reliable information by combining the multiple observations. Although the idea of the processing is new, the current improvement of the basic performance is not so high. In this report, the research trend, the relationship between the conventional methods and the multi-stream approaches, and the mathematical background of the method are discussed

Key words Speech recognition, Multi-Stream

1 はじめに

時間・周波数的に偏在する雑音が重畳された音声を認識する場合，雑音が重畳する時間・周波数領域を認識対象から除くことで，認識性能を向上させることが出来るとする音声に関する missing feature theory の正当性が Lippmann らにより実験的に示されている [1],[2].

背景には，人間の聴覚が周波数毎に独立して機能しているとする考え方があり，明瞭度に関する Fletcher の仮説

$$e_T(c|x^1,x^2) = e_1(c|x^1)e_2(c|x^2)$$

がある [3]．ここで，e_T は2つの周波数帯域を併用した時の誤り率を，e_i は i 番目の帯域のみを利用した時の誤り率を，それぞれ表しており，x^l によって l 番目の帯域の観測信号を表している．すなわちこの仮説は，帯域1と2を併用した時の識別誤り率が，帯域1を利用した時の誤り率と帯域2を利用した時の誤り率との積をして与えられることを仮定している（以下ではこの仮説を error product の仮説と呼ぶ）.

error product の仮定を受け入れるならば，2つの帯域を用いて識別を行った場合の正解識別率 $P_T(c|x^1,x^2)$ は，

$$\alpha(P(c|x^1),P(c|x^2)) = \sum_{l=1}^{2} P(c|x^l) - \prod_{l=1}^{2} P(c|x^l)$$

と計算される．2つの帯域における識別率の幾何平均を，帯域の分割を行わない場合の識別率の目安と考え，α により計算される識別率との差

$$d = \alpha((P(c|x^1),P(c|x^2)) - \sqrt{\prod_{l=1}^{2} P(c|x^l)}$$

を図1に示す．α の値は全ての領域で幾何平均よりも大きく，一方の確率値が小さいほどその差が大きいことが分かる．すなわち複数の帯域の信号を，error product の仮説を満たすように統合することができれば，特定の周波数帯域の音声が著しく劣化していても，その劣化の影響をうけることなく安定して音声の認識を行うことが可能である.

この結果を受け，音声のスペクトルを部分帯域に分割し，雑音の状況に応じて利用する帯域を取捨選択することで高い音声認識性能を得る手法が提案された．さらに部分帯域への分割に止まらず，様々な観点から観測信号を分割し複数の信号系列を利用して音声認識を行う手法が検討されている．[4],[5],[6], [7],[8]

さらに部分周波数帯域に関する error product の仮説を拡大解釈すると，対象に対して多様な観測や特徴抽出を行って入力信号を多元化することで，狭帯域雑音

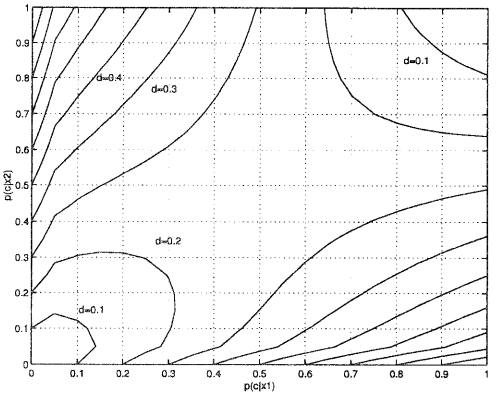

図 1: error product を仮説した識別率と，2つの帯域の認識率の幾何平均との差

などの限定された歪だけでなく，様々な歪に対して頑健な情報の抽出が可能となることが期待される．このような情報の多元化や統合は音声認識において，様々なレベルで行うことができる.

図1には音声認識における，信号の多元化とその統合の段階を示す．一例として，信号波形の多元観測には，空間的に配置された複数のマイクロフォンによる受音，部分周波数帯域毎のパワーや局所スペクトル形状，静的な特徴量と動的な特徴量，複数フレーム長などが考えられる．一方，観測信号の統合方法としては，最も高い確率を与える結果（信号とモデルの組み合わせ）のみを用いる方法や，異なる信号とモデルを用いて認識した結果の重み付け和を確率値の領域で計算する方法などが考えられる.

以下本稿では，multi-stream 型の音声分析・認識の基本原理，従来手法との関係，および研究動向を紹介する.

2 基本アルゴリズム

前節で述べたとおり，multi-stream 型の音声認識では，

- 雑音や誤差の局所化を可能とする帯域の分割や信号の多元観測
- error product の仮説が成り立つような識別器

の両者の実現を前提としている.

雑音の局所化は信号や現象に関する個々の知識を用いることで実現されるが，error product の仮説が成立するような識別器を一般的に構成する方法は知られていない．そのため多くの研究では，雑音の影響を受け

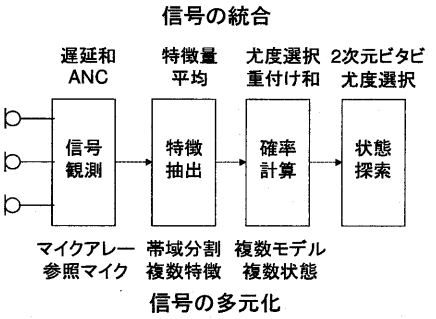

図 2: 音声認識における多元信号観測と統合

ていない信号のみを認識に用いるという方法で識別器を構成している．

図2には筆者らが行った実験の構成を示す．[12] まず 0–4kHz までの帯域を 9 チャネルの部分周波数帯域に分割し，各部分帯域を除去した 9 種類の学習音声を用いて 9 種類の音素 HMM を学習する．次に雑音を重畳した認識対象音声から各部分帯域を除去することで 9 種類の音声を作成し，対応する音素 HMM で音声認識を行う．10 種類（部分帯域を除去した 9 種類の結果と全帯域を用いた結果）のフレーム毎あるいは文毎の確率に従い，認識結果を選択して出力する．

チャープトーンを 0dB の SNR で重畳した音声を認識する実験では，全帯域を認識に用いた場合の認識率が 58% (%correct) であったのに対して，フレーム毎に最大尤度を出力するモデルを動的に選択した場合の認識率は，88%であった．一方，ピアノの独奏を 0dB の SNR で重畳した音声の認識実験では，全帯域を用いた場合が 55%であったのに対して，フレーム毎に最大尤度との尤度差を基準に選択した場合の認識率は 62%であった．一方，モデル毎に得られた 10 種類の認識結果の中から文単位で事後的に選択を行うことで，最大 69%の認識率が得られることが明らかになった．

これらの実験結果から，部分帯域に着目した雑音の居所化が比較的容易に実現できるのに対して，複数の出力結果の効果的な統合が困難であることが分かる．

3 従来の処理手法との関連

多元観測信号の音声認識への直接的な応用法の研究は，現在サブバンド特徴量を使用する方法が主流である．しかし従来の音声信号処理においても頑健性を向上させるために，多元観測信号の統合が間接的に利用されている．以下では，従来の音声処理における多元観測信号の統合処理を概観する．

3.1 アレー信号処理

アレーマイクを用いる選択的受音や，参照マイクを用いるノイズキャンセレーションなどの信号処理は，空間的に分散した多元観測から得られる複数の時間波形を，同じく波形レベルで統合する統合信号処理と見なすことができる．アレー信号処理では，雑音源が空間的に偏在していることが陽に利用され，信号の統合も雑音源と受音器との相対的な位置関係に基づき，固定的に行われることが一般的である．適応的な処理を行う場合の統合基準としては，分散最大化（最小化）などのエネルギー基準が用いられる．

3.2 スペクトルサブトラクション

参照マイクを用いるノイズキャンセレーションが，雑音と音声の空間的な局所性を利用した多元化処理であるのに対して，最も一般的な雑音抑圧処理であるスペクトルサブトラクションは，時間領域での雑音の局所性を利用した多元化処理と考えることができる．雑音のスペクトルの推定は音声の休止区間に行われ，雑音重畳音声と推定雑音の 2 つのスペクトル信号は，パワースペクトル領域での減算を介して統合される．

3.3 動的特徴

静的なスペクトル特徴と動的なスペクトル特徴の併用は，特徴量の多元化の最も直接的な実現手法である．動的特徴と静的特徴の統合は，両者が独立であることを仮定して，確率計算のレベルで固定的に行われるのが一般的である．この時，識別正解率は，

$$P(\hat{c}|\mathbf{x}^{\mathsf{s}})P(\hat{c}|\mathbf{x}^{\mathsf{d}})$$

ただし，

$$\hat{c} = \mathbf{argmax}P(c|\mathbf{x}^{\mathsf{s}})P(c|\mathbf{x}^{\mathsf{d}})$$

で与えられるため，どちらか一方の特徴量の観測にのみ誤差や歪が偏在する場合でも，最終的な認識性能の劣化は避けられない．

3.4 混合ガウス分布

混合ガウス分布を利用することで，単一の観測信号に対して複数の確率（密度）値が状態毎に計算されることは，確率レベルでの信号の多元化と見なすことが

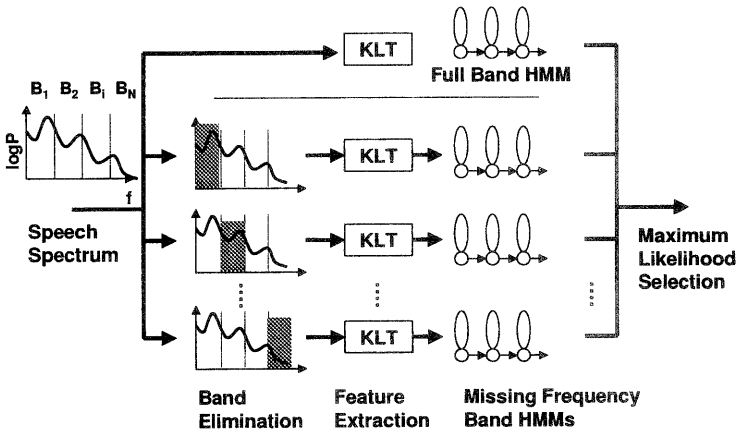

図 3: 帯域除去モデルによるマルチストリーム音声認識

できる．通常の認識器における信号の統合は，確率値の重み付け和として与えられるが，重みは固定的に用いられることが一般的であり，統合基準に基づく動的な統合処理が機能している訳ではない．混合分布による多元化では，状態 s_i における出力確率が，

$$\sum_{j=1}^{M} w_j f(x|\Theta_{ij})$$

により与えられるため，一部の分布パラメータ Θ_{ij} の推定に誤差や誤りがあっても，その影響が全体に及ぶことは少ないと考えられる．

3.5 PMC 法，複数モデル法

PMCや2次元ビタビと呼ばれる手法は，状態レベルでの多元化と考えることができる．通常のデコーダにおいては単一のHMM状態で表現される音素状態を，重畳雑音の状態などの発話内容とは独立な情報に対応させて多元化する．多元化された状態の統合は，通常のHMMと同様に最も高い累積尤度を与える状態系列を選択することで実現される．

このような多元化された状態の極端なケースとして，(男性モデルと女性モデルといった) 複数モデルセットを用いて独立にデコーディングを行った後に，高い累積尤度を出力したモデルに対応する結果を選択・出力する，複数モデルとその尤度選択に基づく手法がある．話者クラスタリングに基づいて作成された複数のモデルセットを用いる高速話者適応法も，同様に複数モデルの選択利用に基づく方法であるが，選択基準には認識結果に対する尤度ではなく，確率的に定義された話者間測度が用いられることが一般的である．

4 multi-stream 手法の研究動向

本年 10 月に行われた ICSLP (Internationa Conference on Spoken Language Processing) では，multi-stream に関連して非常に多くの研究発表が報告された．また 5 月に行われた ICASSP (IEEE International Conference on Acoustic Speech and Signal Procesing) においても multi-stream 型の音声認識に関する発表が数件見られた．以下本節では，これらの国際会議における multi-stream 型音声認識の研究報告の主たる論点について概説する．

- stream（特徴量）の構成方法
  多くの報告で，部分周波数帯域への分割に基づき stream が構成されている．音声認識特徴量としては，部分帯域内の対数パワースペクトルを DCT 分析したサブバンドケプストラム係数が多く用いられている．全帯域をブロック DCT して得られる単一の特徴ベクトルを用いる方法，[14] や，複数の帯域分割方法を併用する手法 [15] なども報告さ

れている．音素毎に異なる識別関数を利用する形式の多元化も報告されている．[17]

- stream の選択・重み基準

  多元化された信号系列から，適切な信号系列を選択する選択基準は，確率的な基準と音響的な基準の２つに大別される．多くの選択型のシステムでは，尤度に基づいて信号系列を選択するが，部分周波数帯域の音響的な性質を利用することも有効である．

  さらに特定 stream を選択的に使うのではなく，stream 毎に異なる重み α を用いて，

$$f(x|c) \approx \prod_{i=1}^{i=N} f(x_i|c)^{\alpha_i}$$

  のように確率（密度）計算レベルで統合して用いる方法も多く試みられている．重み α_i は，音響的な基準から求める方法 [19], [20], や相互情報量から求める方法 [21]，パワースペクトルの時間周波数パタンから重みを求める方法 [23] 等が報告されている．また，予め用意された学習データを用いて重みパラメータを識別学習する方法も多く検討されている．[25]

- stream 間の同期

  確率値の統合を行う場合，状態遷移の stream 間の同期に関する議論がある．状態間の遷移ネットワークをマルコフ場でモデル化し，同期を取ることで得られる性能改善に関する報告があるのに対して，特徴量毎にことなる状態遷移を許すことで得られる性能の改善に関しても報告がなされている．[27],[28],[29],[30],[31]

5 むすび

多様な観測を行いその結果を統合することで頑健な音声分析を行う，新しい音声分析の方向を紹介した．当該手法の基本的な考え方は合理的であり，これまでの様々な音声分析処理を包含しうるものである．しかしその実現方法，特に情報の統合方法に関しては未だ十分に検討がされておらず，報告されている様々な実験結果においてもいまだ性能のばらつきが大きい．

今後も音声認識を中心にした，音声言語インタフェース技術は高度化を続けると考えられる．ユーザや外界の状況を柔軟に理解しうる知的システムを実現するためには，音声だけでなく多様なメディア信号の統合メカニズムを高度化する必要があり，頑健なメディア信号処理法としての，多元信号とその統合に関する研究の進展に大きな期待が寄せられている．

謝辞

日頃ご議論いただく，名古屋大学統合音響情報研究拠点の教官・研究員・学生諸氏に感謝いたします．

参考文献

[1] R.Lippmann and B.Carlson "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise", Proc. of European Conference on Speech Communication and Technology.(Eurospeech '97), KN37–40, 1997

[2] R.Lippmann and B.Carlson "Robust speech recognition with time-varying filtering, interruptions, and noise." Proc. of 1997 IEEE Workshop on Automatic Speech Recognition and Unederstanding, pp. 365–372

[3] 三浦種敏 監修：新版 聴覚と音声 電子情報通信学会編

[4] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands." Proc. of International Conference on Spoken Language Processing.(ICSLP'96), pp.426–429, Philadelphia, October 1996.

[5] H. Hermansky, S. Tibrewala, and M. Pavel. "Towards ASR on partially corrupted speech." Proc. of International Conference on Spoken Language Processing.(ICSLP96), pp.1579–1582, October 1996.

[6] H. Bourlard and S. Dupont, "Subband-based speech recognition", Proc. of IEEE International Conference on Acoustic Speech and Signal Processing.(ICASSP'97), pp. 2 125–128, 1997

[7] H. Bourland, "Non-stationary multi-channel (multistream) processing towards robust and adaptive ASR", Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp.1–10, 1999

[8] A. Janin, D. Ellis, and N. Morgan "Multi-stream speech recognition: Ready for prime time?" Proc. of European Conference on Speech Communication and Technology.(EUROSPEECH'99) pp.591–594, 1999

[9] S.Okawa, E.Bocchieri and A.Potamianos, "Multi-band speech recognition in noisy environments," Proc. of IEEE International Conference on Acoustic Speech and Signal Processing.(ICASSP'98), pp.641–644

[10] H. Nock and S.Young, "Loosely coupled HMMs for ASR," Proc. of International Conference on Spoken Language Processing.(ICSLP 2000), 2000

[11] J.Fiscus "A post-processing system to yield reduced word error rates: recogniser output voting error reduction (ROVER)", Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'97), pp. 347–352, 1997

[12] 河村良尊，武田一哉，板倉文忠：『特定帯域不使用モデルを用いた雑音環境下の音声認識』音講論集 2-Q-16, pp.I 121–122 2000.9

[13] N. Mirghafori and N. Morgan, "Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers", Proc. of IEEE International Conference on Acoustic Speech and Signal Processing.(ICASSP'98), 1998

[14] K.Paliwal and J. Chen "Robust feature extraction for speech recognition." Proc. of The Seventh Western Pacific Regional Acoustics Conference (WESTPRAC VII), pp. 61–66, 2000

[15] P. McCourt, N. Harte and S. Vaseghi, "Combined temporal and spectral multi-resolution phonetic modelling," Proc. of European Conference on Speech Communication and Technology.(EUROSPEECH'99) pp. III 1111–1114, 1999

[16] A. Hagen and H. Bourlard, "Using multiple time scales in the framework of multi-stream speech recognition," Proc. of International Conference on Spoken Language Processing.(ICSLP 2000), 2000

[17] A. Halberstadt and J. Glass, "Heterogeneous Measurements and Multiple Classifiers for Speech Recognition," Proc. of IEEE International Conference on Acoustic Speech and Signal Processing.(ICASSP'98) 1998

[18] J. Ming, P. Jancovic, P. Hanna, D. Stewart and F. Smith, "Robust feature selection using probabilistic union models", Proc. of International Conference on Spoken Language Processing.(ICSLP 2000), 2000

[19] F. Berthommier, H. Glotin, E. Tessier and H. Bourlard, "Interfacing of CASA and partial recognition based on a multistream technique," Proc. of IEEE International Conference on Acoustic Speech and Signal Processing.98 (ICASSP'98) 1998

[20] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," Proc. of IEEE International Conference on Acoustic Speech and Signal Processing.(ICASSP'97), pp. 2 1255–1258, 1997

[21] S. Okawa, T. Nakajima and K. Shirai, "A recombination strategy for multi-band speech recognition based on mutual information criterion," Proc. of European Conference on Speech Communication and Technology.(EUROSPEECH,99) pp. II 603–606 1999

[22] H. Glotin and F. Berthommier, "Test of several external posterior weighting functions for multiband Full Combination ASR", Proc. of International Conference on Spoken Language Processing.(ICSLP 2000), 2000

[23] B. Raj, M. Seltzer and R. Stern, "Reconstruction of damaged spectrographic features for robust speech recognition" Proc. of International Conference on Spoken Language Processing.(ICSLP 2000), 2000

[24] J. Barker, L. Josifovski, M. Cooke and P. Green, "Soft decisions in missing data technoques for robust automatic speech recognition," Proc. of International Conference on Spoken Language Processing.(ICSLP 2000), 2000

[25] C. Christophe, H. Jean-Paul and F. Dominique, "Towards a global optimization scheme for multi-band speech recognition", Proc. of European Conference on Speech Communication and Technology.99 (EUROSPEECH'99), pp. II 587-590, 1999

[26] P. McMahon, P. McCourt, S. Vaseghi, "Discriminative weighting of multi-resolution sub-band cepstral features for speech recognition", Proc. of International Conference on Spoken Language Processing.(ICSLP'98),

[27] K. Daoudi, D. Fohr and C. Antoine "A new approach for multi-band speech recognition based on probabilistic graphical models", Proc. of International Conference on Spoken Language Processing.(ICSLP 2000) 2000

[28] G. Gravier, M. Sigelle and G. Chollet, "A Markov random field based multi-band model," Proc. of IEEE International Conference on Acoustic Speech and Signal Processing.(ICASSP2000), pp.III 1619–1622, 2000

[29] S. Matsuda, M. Nakai, H.Shimodaira and S.Sagayama, "Asynchronous-transition HMM", Proc. of IEEE International Conference on Acoustic Speech and Signal Processing.(ICASSP 2000), pp.II 1005–1008, 2000

[30] C. Cerisara, D. Fohr and J. Haton, "Asynchrony in multi-band speech recognition," Proc. of IEEE International Conference on Acoustic Speech and Signal Processing.(ICASSP2000), pp. II 1121–1124, 2000

[31] N. Mirghafori and N. Morgan, "Sooner or later: exploring asynchrony in multi-band speech recognition" Proc. of European Conference on Speech Communication and Technology.(Eurospeech ' 99) Budapest, 1999.