社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69（７）
2007／12／20

探索空間のエントロピーに基づく特徴量ストリームの動的な統合

佐藤 庄衛[†,††] 奥 貴裕[†] 本間 真一[†] 小林 彰夫[†] 今井 亨[†]
都木 徹[†] 小林 哲則[††]

† NHK 放送技術研究所 〒157–8510 東京都世田谷区砧 1–10–11
†† 早稲田大学理工学研究科 〒169–8555 東京都新宿区大久保 3–4–1
E-mail: †{satou.s-gu,oku.t-le,homma.s-fc,kobayashi.a-fs,imai.t-mq,takagi.t-fo}@nhk.or.jp, ††koba@waseda.jp

あらまし 本報告では，頑健な音声認識を実現するため，異なる音響的な特徴を表す複数のストリームから得られる音響尤度を，動的に統合する手法を提案する．提案法では，各ストリームの観測による探索空間のエントロピーの削減量に応じた，動的なストリーム重みを導入した．ニュース音声の認識実験において，エネルギーと共振周波数の推移量（AM，FM）に基づく三種類の特徴量ストリームを提案法によって統合した結果，単一ストリームで得られる単語誤認識率に比べ，雑音の多い現場リポート部分で 9.2%，自由発話が含まれる対談部分で 4.7%の単語誤り削減率が得られた．

キーワード 音声認識，ストリーム統合，エントロピー

Dynamic Integration of Multiple Feature Streams Utilizing Entropy of Search Hypotheses

Shoei SATO[†,††], Takahiro OKU[†], Shinich HOMMA[†], Akio KOBAYASHI[†], Toru IMAI[†], Tohru TAKAGI[†], and Tetsunori KOBAYASHI[††]

† NHK Science and Technical Research Laboratories 1–10–11 Kinuta, Setagaya-ku, Tokyo, 157–8510, Japan
†† Faculty of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169–8555, Japan
E-mail: †{satou.s-gu,oku.t-le,homma.s-fc,kobayashi.a-fs,imai.t-mq,takagi.t-fo}@nhk.or.jp, ††koba@waseda.jp

Abstract We present a novel method of integrating the likelihoods of multiple feature streams, representing different acoustic aspects, for robust speech recognition. The proposed integration algorithm dynamically calculates a frame-wise stream weight based on an entropy reduction caused by observing an input feature. Japanese broadcast news recognition experiments, integrating feature streams representing energy, amplitude drifts, and resonant frequency drifts, showed that the proposed method reduced error words by 9.2% in field reports and 4.7% in spontaneous commentaries relative to the best result obtained from a single stream.

Key words Speech recognition, Stream integration, Entropy

1. はじめに

近年の統計モデルに基づいた音声認識は，大量の学習データが利用できるようになったため，様々な実用的アプリケーションに応用されるようになってきた．例えば，NHK では，生放送番組への字幕付与を行うために，音声認識を利用している [1]．

一方，音声認識を利用するにあたり，多くのアプリケーションに共通する問題点として，背景雑音や自由な発話スタイルによる誤認識の増加があげられる．このような誤認識の増加は，放送音声の認識においても，字幕制作などの実用化を拡大するための課題となっており，“現場リポート” や “対談” 部分の認識精度の向上が望まれている．

人間は，様々な音響的手がかりを柔軟に利用することで，背景雑音や発話スタイルに頑健な音声認識を実現していると考えられる．そこで筆者らは，異なる音響的な特徴を表す複数のストリームから得られる音響尤度を，各ストリームの観測によって得られる探索空間のエントロピー削減量に応じた重みに基づき，動的に統合する手法を提案してきた [2], [3]．本稿では，提

案手法の効果を，様々な条件で詳細に調べたので，その結果を報告する．

これまでも，複数のストリームの統合手法が数多く提案されてきているが，これらの手法は次の三種類の方法に分類される．第一の手法では，変換行列を用いて，複数の特徴量を識別的に一つのストリームに統合する．これらの変換行列は，PCA, LDA, HLDA など，分析的に最適化される [4], [5]．

第二の手法では，各ストリームから得られる音響尤度を，静的に統合する．この手法では，各 HMM の状態の尤度は，各ストリームで得られる対数尤度の重みづけ和となる．このストリーム重みは，学習データを用いて事前に最適化することができ，この最適化の基準には，ML, MMI, MCE, ME などの基準が用いられる [6]〜[8]．

第三の手法では，上述の尤度統合重みを，入力音声に応じて動的に制御する．このような動的な手法では，ストリームの識別能力を用いて重みを制御するものが多く，[9] では上位に残った探索仮説 (n-best) の尤度比を用いてストリームの識別能力を評価し，読み上げ数字の認識精度の向上を報告している．しかし，大語彙連続音声認識では，音素環境ごとに学習された同一音素のモデルが多数存在するため、少数の仮説で評価された識別能力は，過小な評価になりやすい．一方，全ての状態を探索空間として求めたエントロピーを用いて，動的なストリーム重みを与える手法も提案されている．[10] では，読み上げ数字の認識タスクにおいて，数 100 規模の HMM の状態全ての尤度から，探索空間のエントロピーを求め，動的なストリーム重みを与えた．しかし，リアルタイムで動作する大語彙連続音声認識では，全ての HMM の状態の尤度を算出するのは計算量の観点から困難である．

提案法は，このエントロピーに基づく動的ストリーム重みを，リアルタイム大語彙連続音声認識に適用したものである．本手法では，尤度計算を要する状態数を増やすことなく識別能力を評価するため，探索パス中でアクティブな状態のみを探索空間とし，エントロピーを求める．さらに，時刻ごとに異なる探索空間の周辺エントロピーを考慮し，周辺エントロピーとストリームを観測した後の探索空間のエントロピーから，エントロピーの削減量に比例したストリーム重みを与えた．これにより，入力音声の背景雑音や発話スタイルに頑健なストリームほど，大きな重みづけを施した探索が期待される．

本稿では，AM, FM などの変調情報が聴覚で抽出されているという知見に基づき，聴覚フィルタの出力から，これらの変調情報を含む三種類のストリームを統合し，ニュース音声の認識実験を行った．

2. 動的なストリーム尤度の統合

提案する動的ストリーム尤度の統合手法の概要を，図 1 に示す．提案法では，様々な特徴量ストリームを統合しうるが，本稿では後述の "*Energy*", "*AM*", "*FM*" の三種類のストリームを統合する．提案法では，n 番目のフレームの特徴量 $x_k(n)$ (k はストリーム番号) が入力された時点でアクティブな探索仮説 (HMM ノード) を探索空間 $\Lambda(n)$ としたときの，$\Lambda(n)$ の周辺

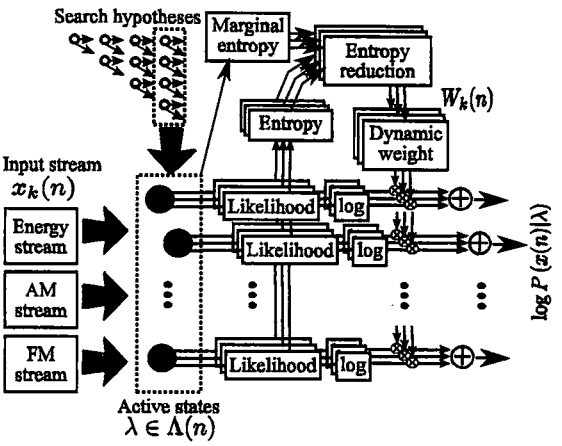

図 1　提案するストリーム尤度の動的統合の概要．

エントロピーと，ストリーム観測後の状態 $\lambda \in \Lambda(n)$ の尤度から求められる $\Lambda(n)$ のエントロピーの差に基づいて，ストリーム $k \in \{1 \cdots K\}$ (K は統合するストリームの総数) の統合重み $W_k(n)$ を動的に決める．この動的な重みを用いて，$\lambda \in \Lambda(n)$ の識別能力が高いストリームに大きな重みを与えて，各ストリームの尤度 $P(x_k(n)|\lambda)$，$k \in \{1 \cdots K\}$ を統合し，探索を行う．

2.1 エントロピーの削減量に応じたストリーム重み

提案法では，ストリーム $x_k(n)$ を観測する事によって得られる探索空間 $\Lambda(n)$ のエントロピーの削減量 $I_k(n)$ に基づいて，ストリームの識別能力を評価する．$x_k(n)$ を観測することによって得られるエントロピーの削減量は，

$$I_k(n) = H^0(n) - H_k(n) \tag{1}$$

で与えられる．ここで，$H^0(n)$ は $\Lambda(n)$ の周辺エントロピー，つまり特徴量を観測する前の $\Lambda(n)$ のエントロピーである．$H_k(n)$ は $x_k(n)$ を観測した後の $\Lambda(n)$ のエントロピーであり，$P(x_k(n)|\lambda)$ から求められる．

一方，式 (1) はストリーム k のエントロピー $H_k(n)$ を，傾きが -1 の線形変換とオフセット $H^0(n)$ を用いて，ストリーム識別能力を表す非負の量に変換しているととらえる事ができる．提案法のように，アクティブな HMM ノードのみのエントロピーを考える場合，$H_k(n)$ のとりうる値は入力フレーム n ごとに異なるため，事前に設定した固定値のオフセットで $I_k(n)$ の非負を保証できない．そこで，本稿では，非負の識別能力を保証するため，$\lambda \in \Lambda(n)$ の事前分布を一様分布として，$H^0(n)$ はアクティブなノードの個数 $L(n)$ にのみ依存し，次式により与えられるものとする．

$$H^0(n) = \log L(n) \tag{2}$$

$x_k(n)$ 観測後の $\Lambda(n)$ のエントロピーは，次式で与えられる．

$$H_k(n) = -\sum_{\lambda \in \Lambda(n)} P(\lambda|x_k(n)) \log P(\lambda|x_k(n)) \tag{3}$$

λ の事後確率 $P(\lambda|x_k(n))$ は，$P(\lambda_1) = \cdots = P(\lambda_{L(n)}) =$

$1/L(n)$，および $P(x_k(n)) = 1/L(n)\sum_{\lambda\in\Lambda(n)}\check{P}(x_k(n)|\lambda)$ を仮定し，次式により与えられる．

$$P(\lambda|x_k(n)) = \frac{\check{P}(x_k(n)|\lambda)}{\sum_{\lambda_j\in\Lambda(n)}\check{P}(x_k(n)|\lambda)} \tag{4}$$

一般的に，尤度を統合しようとしているストリームの識別能力に差異があるのに加え，式 (4) では尤度の絶対値が考慮されないため，事後確率の信頼度もストリームにより異なる．そこで，信頼度が低く認識精度の低下を引き起こすストリームの音響尤度への寄与率を補正するため，

$$\check{P}(x_k(n)|\lambda) = P(x_k(n)|\lambda)^{w_k} \tag{5}$$

を導入した．ここで，w_k はストリーム k の信頼度を補正する静的な重みである．この静的な重みは，学習データを用いて ME や MCE 基準で最適化できる．

提案法では，この補正尤度を対数領域で動的な重み $W_k(n)$ を与えて統合する．

$$\log P(x_k(n)|\lambda) = \sum_{k=1}^{K} W_k(n)\log\check{P}(x_k(n)|\lambda) \tag{6}$$

最終的に，次式の再帰計算によりビタビ探索が行われる．

$$\alpha_i(n) = \max_j \alpha_j(n-1)a_{ji}P(x(n)|\lambda_j) \tag{7}$$

ここで，$\alpha_i(n)$ はノード λ_i の累積ビタビスコアであり，a_{ji} はノード j から i への遷移確率である．

本稿では，エントロピー $H_k(n)$ のみを用いた二つの従来法 $W_k^{MinH}(n)$，W_k^{InvH} と，提案する動的重み $W_k^{MI}(n)$ を比較する．

2.2 従来法（選択的重み ($MinH$)）

第一の従来法は，最もエントロピーが小さいストリームを選択的に用いる方法であり [10]，ストリーム統合重み W_k^{MinH} は次式により得られる．

$$W_k^{MinH}(n) = \begin{cases} 1.0 & \text{if } k = \arg_k \min H_k(n) \\ 0.0 & \text{otherwise} \end{cases} \tag{8}$$

2.3 従来法（逆数を用いた重み ($invH$)）

第二の従来法は，エントロピーの逆数を用いて，情報量の大きなストリームに大きな統合重みをあたえる方法であり [10]，統合重み $W_k^{invH}(n)$ は次式により得られる．

$$W_k^{invH}(n) = \frac{1/H_k(n)}{\sum_{j=1}^{K} 1/H_j(n)} \tag{9}$$

2.4 提案法（エントロピー削減量を用いた重み）(MI)

提案法では，式 (1) によるエントロピーの削減量を用いて統合重みを与える．この方法では，前述の二つの方法と異なり，周辺エントロピー $H^0(n)$ が考慮されている．上述の $W_k^{invH}(n)$ と同様に，統合重み $W_k^{MI}(n)$ は $I_k(n)$ を正規化して，次式により得られる．

$$W_k^{MI}(n) = \frac{I_k(n)}{\sum_{j=1}^{K} I_j(n)} \tag{10}$$

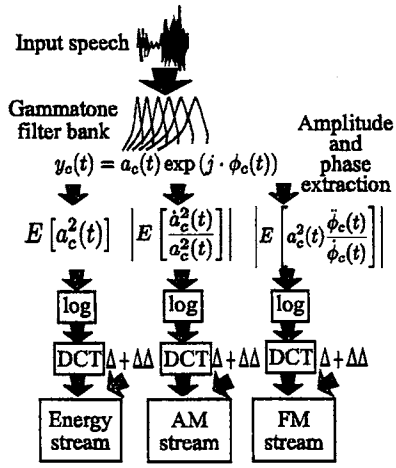

図 2　統合した三種のストリーム

3. 統合したストリーム

提案法の有効性を調べるため，図 2 に示す三種の特徴量を統合した．直交基底を用いた複素ガンマトーンフィルタバンク [11] の出力 $y_c(t)$ から振幅 $a_c(t)$ と位相 $\phi_c(t)$ 成分を得た後，後述の三種のフレーム内の期待値を得る．ここで，t は量子化サンプル番号，c はフィルタバンク番号である．特徴量は，この三種の期待値の対数の DCT 係数から求めた 12 次のケプストラム（C0 をのぞく）と対数エネルギーおよびそれらの一次と二次の回帰係数の 39 次元からなるストリームである．

3.1 Energy

第一のストリームは，周波数エネルギー分布を表すものであり，MFCC と同等の特徴量である．フィルタ c に対応する要素 $EG_c(n)$ は，次式より得られる．

$$EG_c(n) = E_n\left[a_c^2(t)\right] = \sum_{t=1}^{T^w} a_c^2\left(T^s\cdot n + t\right) \tag{11}$$

ここで，T^w は分析フレーム幅，T^s はフレームシフト幅である．

3.2 AM

第二のストリームは，フィルタ出力エネルギーの時間変動量に着目した特徴量である．ここでは，知覚検知限 (JND) に関するウエーバーの法則を考慮し，エネルギーの変動量 $\dot{a}_c^2(t) = a_c^2(t) - a_c^2(t-1)$ とエネルギー $a_c^2(t)$ の比を特徴量とした．

$$AM_c(n) = \left|E_n\left[\frac{\dot{a}_c^2(t)}{a_c^2(t)}\right]\right| = \left|\sum_{t=1}^{T^w}\frac{\dot{a}_c^2(T^s\cdot n+t)}{a_c^2(T^2\cdot n+t)}\right| \tag{12}$$

3.3 FM

第三のストリームは，フィルタ出力の共振周波数の変動量に着目した特徴量である．ここでは，位相 $\phi(t)$ を微分することで得られる瞬時共振周波数 $\dot{\phi}(t) = \phi(t) - \phi(t-1)$ と共振周波数の変動量 $\ddot{\phi}(t) = \dot{\phi}(t) - \dot{\phi}(t-1)$ を用いて，JND を考慮した上で $\ddot{\phi}(t)/\dot{\phi}(t)$ を特徴量とした．さらに，入力音声が無く，信号のエネルギーが小さい部分の影響を小さくするため，エネル

表 1 単独ストリームの WER.

Feature stream	WER [%] all	field	spon.
Energy [39]	7.4(6.5)	6.7(7.6)	12.7(11.1)
AM [39]	10.9(9.3)	10.6(9.2)	18.2(15.5)
FM [39]	11.6(9.8)	12.0(10.3)	18.5(15.4)

[] 内はストリームの次元数.
() 内は挿入誤りを除いて評価した WER.

表 2 連結特徴量での WER.

Feature-stream	WER [%] all	field	spon.
EAF [111]	8.9(7.4)	9.0(7.8)	14.8(12.5)
HLDA [39]	8.2(7.1)	8.8(7.8)	12.3(10.6)
HLDA [57]	7.5(6.5)	7.8(6.9)	12.3(10.3)

[] 内はストリームの次元数.
() 内は挿入誤りを除いて評価した WER.

ギーで重みづけをして特徴量とした.

$$FM_c(n) = \left| E_n \left[a_c^2(t) \frac{\ddot{\phi}_c(t)}{\dot{\phi}_c(t)} \right] \right| = \left| \sum_{t=1}^{T^w} a_c^2(T^s n + t) \frac{\ddot{\phi}_c(T^s n + t)}{\dot{\phi}_c(T^s n + t)} \right| \quad (13)$$

4. 認識実験

4.1 実験条件

提案するストリーム統合手法による認識精度の改善を確認するため，NHK が 2004 年 7 月に放送したニュース番組のうち，認識率の低下が見られた 16 のニュース項目を評価音声とした．評価音声は，643 発話の不特定男性発話，8,391 単語 ("*all*") であり，現場リポート音声 395 発話，4,905 単語 ("*field*") と，対談調音声 231 発話，3,220 単語 ("*spon.*") が含まれる．うち，97 発話は現場リポートと対談調音声の両方に含まれ，この両者に属さない 114 発話はアナウンサーによるスタジオでの原稿読み上げ音声であった．

本実験で用いる各ストリーム用に，4,000 共有状態，16 混合分布，5 状態 3 ループのトライフォン HMM を 2004 年 6 月までに放送された NHK のニュース番組の不特定男性発話 100 時間分から学習した．特徴量は，40 チャンネルのガンマトーンフィルタバンクの出力を，フレーム幅 25ms，フレームシフト幅 10ms で切り出した後に抽出された．それぞれの回帰係数を含む特徴量は，学習データに対して平均が 0，分散が 1 となるように正規化して用いた．

認識に用いた言語モデルは，過去 10 年分のニュース原稿 (127M 単語) を元に，直近に入力された原稿に n-gram カウントレベルで重みをつけて学習した適応化言語モデル [12] (語彙 61K) である．評価音声に対するトライグラムパープレキシティーは 21.9，未知語率は 0.4%であった．

さらに，異なるダイナミックレンジを示す種々の尤度統合手法の認識結果を比較するため，音響スコアのスケーリングを行った．このスケーリング係数は，探索ビーム幅を 200，アクティブノード数の最大値 2,000，言語スコア重み 10 という条件で WER が最小になるように調整した．このスケーリング係数は，本実験をとおして 0.4〜2.5 の値に調整され，各フレームでアクティブなノード数の平均値が，2,142〜2,425 であることから，ほぼ同等の探索条件であることが確認された．

4.2 単独ストリームの性能

三種のストリームを単独に用いた場合の単語誤認識率 (WER) を，表 1 に示した．また，挿入誤りを評価しなかった場合の WER を表中の () 内に示した．ここで，挿入誤りを除外した WER を併記したのは，評価音声に多く含まれる言いよどみ部分を正当に評価する事が難しかったためである．これらの言いよどみの多くは，複数の単語で構成されており，単語の一部に不明瞭な音素や音素の欠落がある．このような部分は，正当な正解を与える事が難しく，字幕制作用の大語彙連続音声認識 [1] では，後段の修正装置で削除される部分である．さらに，これらの言いよどみの削除にかかるコストは，非常に小さい．評価に用いた正解単語列からも，言いよどみは除外した．

三種のストリーム単独での認識結果を比較すると，変調特徴量 ("*AM*", "*FM*") では，"*Energy*" ストリームに比べて WER が増加した．

三種のストリームを連結した特徴量での認識結果を，表 2 に示した．表中の "*EAF*" は，三種のケプストラム係数と対数エネルギー ($12 \times 3 + 1$) と，それらの回帰係数 (37×3) の 111 次元のストリームを用いた場合の認識結果である．"*HLDA*" は，この 111 次元のストリームを，HLDA-MLLT 行列によって，39 次元もしくは 57 次元に圧縮・無相関化したストリームでの認識結果である．"*HLDA*(57 次元)" は，評価音声の WER を基準にフィボナッチ探索 [13] で最適な次元数を求めた場合の WER である．

ここで，"*EAF*" は "*Energy*" と比較して WER の増加が見られた．さらに，"*HLDA*" によって識別的に変換したストリームにおいても，"*Energy*" と比較すると，"*spon.*" の条件でしか WER の改善は見られなかった．

4.3 ストリームの統合結果

ストリームの統合実験で用いる HMM は，全て同一の状態共有構造と状態遷移確率を有している．この状態共有構造と状態遷移確率は，上述の "*EAF*" と同じである．統合用の HMM は，各状態の状態出力確率を与える混合分布パラメータのみを学習した．この共有構造と遷移確率の共有により，デコーダーでは単一の木構造探索ネットワーク上で，ストリーム尤度を動的に変化させながら探索できる．

ストリームを統合した場合の WER を，表 3 に示した．この実験では，ストリームの信頼度を補正する静的な重み w_k 全てに，単位重み 1.0 を与え，ストリームの信頼度を補正しなかった．表中の "*Static*" は，動的なストリーム重み $W_k(n)$ の全てに単位重み 1.0 を与え，静的にストリームを統合した場合の結果である．"*Static*" の WER は，"*Energy*" で得られた WER よりも大きかった．"*MinH*", "*InvH*", "*MI*" は，式 (8)，式 (9)，式 (10) で与えられる動的なストリーム重みを用いた場

表 3 ストリーム尤度を統合した場合の WER．静的な尤度重み w_k を用いない場合．($\forall k,\ w_k = 1.0$).

Integration	WER [%]		
$W_k(t)$	all	field	spon.
Static	7.9(6.6)	7.9(6.8)	13.7(11.4)
MinH	7.5(6.3)	7.2(6.3)	13.0(11.0)
InvH	7.6(6.3)	7.5(6.5)	13.0(11.0)
MI	7.8(6.4)	7.6(6.5)	13.5(11.1)

() 内は挿入誤りを除いて評価した WER.

表 4 ストリーム尤度を統合した場合の WER．最適化された静的な尤度重み w_k を用いた場合．$\{w_k\} = \{0.51(Energy), 0.28(AM), 0.21(FM)\}$.

Integration	WER [%]		
$W_k(t)$	all	field	spon.
Static	7.6(6.5)	7.5(6.5)	13.3(11.3)
MinH	7.3(6.2)	7.0(6.1)	12.2(10.5)
InvH	7.5(6.3)	7.4(6.4)	12.5(10.6)
MI	7.0(5.9)	6.9(6.0)	12.1(10.3)

() 内は挿入誤りを除いて評価した WER.

合の認識結果である．これらの認識結果においても，"*all*" では "*Energy*" に比べて WER の削減は見られなかった．しかし，"*MinH*" では "*field*" で WER の削減がみられ，挿入誤りを除いて評価すると，全ての動的統合法で "*field*" の改善が見られた．これらの結果は，"*HLDA*(57 次元)" で "*spon.*" だけが改善した結果と対照的な結果である．

この実験で，"*MinH*" が "*Energy*"，"*AM*"，"*FM*" の各ストリームを選択した割合は，それぞれ 70%，16%，15%であった．同様に，"*InvH*" から得られた重みの平均の比は，"*Energy*"，"*AM*"，"*FM*" それぞれに対して，0.45:0.28:0.27 であり，"*MI*" から得られた重みの平均の比は，0.37:0.31:0.32 であった．この比の差を見ると，"*MI*" ではストリームの統合重みの差が小さいため，"*InvH*" に比べて WER が増加したと考えられる．

静的な尤度重み w_k を最適化した場合の WER を，表 4 に示した．約 9 時間分の学習データから，最大エントロピー基準[8] を用いて w_k を最適化し，$\{w_k\} = \{0.51(Energy), 0.28(AM), 0.21(FM)\}$ が得られた．

この実験で，"*MinH*" が "*Energy*"，"*AM*"，"*FM*" の各ストリームを選択した割合は，それぞれ 99.8%, 0.2%, 0.1%であり，その他の動的重み付け法により得られた評価音声での各ストリームの平均重みは，"*InvH*" で 0.49:0.26:0.24，"*IEWAT*" で 0.99:0.008:0.001，"*MI*" で 0.61:0.23:0.15 であった．

この静的重み w_k の最適化により，提案法 "*MI*" の WER が "*field*"，"*spon.*" の両条件で大きく改善された．この結果は，"*HLDA*" が "*spon.*" のみを，"*Static*" が "*field*" のみを改善したという結果と対照的である．

最終的に，エントロピーの削減量に基づく提案法 "*MI*" では，最良の結果が得られた従来法 "*MinH*" の WER(7.3%) に対して，4.1%の WER 削減率が得られた．また，"*Energy*" の WER と比較すると，"*all*" で 5.4%，"*field*" で 9.2%，"*spon.*" で 4.7% の WER 削減率が得られた．

表 5 音響モデル中の全ての状態を探索空間とした場合の認識結果．

Integration	WER [%]		
$W_k(t)$	all	field	spon.
MinH.full	7.3(6.2)	7.0(6.1)	12.2(10.5)
InvH.full	7.4(6.2)	7.2(6.3)	12.6(10.6)
MI.full	7.0(5.9)	6.8(5.9)	12.2(10.3)

() 内は挿入誤りを除いて評価した WER.

さらに，提案法 "*MI*" および "*MinH*" と "*Energy*" の認識結果の有意差を，一対の標本による平均の t 検定 [14] を用いて調べると，有意水準 5%で "*MI*" は有意であるが "*MinH*" は有意ではなかった．また，"*MI* 7.0% (5.9%)" および "*MinH* 7.3% (6.2%)" の有意差を調べたところ，挿入誤りを考慮した WER では有意ではなかったが，挿入誤りを除いた評価では有意な差が見られた．

4.4 静的重みの必要性

提案法による動的重み $W_k^{MI}(n)$ での静的な重みの必要性を調べるため，静的重み w_k を用いないで算出した $W_k^{MI}(n)$ を用いて，最適化した静的重みを用いて計算した音響尤度 $\tilde{P}(x_k(n)|\lambda)$ を統合する実験を行った．その結果，表 4("*MI*") に示した WER は，"*all*" で 7.8%，"*field*" で 7.3%，"*spon.*" で 13.7%に増加した．この結果より，ストリーム統合重みの推定の際にも，静的重み w_k による補正を行った尤度 $\tilde{P}(x_k(n)|\lambda)$ から，動的重みを求める必要がある事が示された．

4.5 探索空間の近似の妥当性

提案法では，大語彙連続音声認識に適用するため，エントロピーを求める探索空間として，アクティブな状態を用いた．この近似の妥当性を調べるため，音響モデルに学習された全ての状態（3,972 状態）のエントロピーを用いてストリームの統合を行った．表 5 は，3,972 状態のエントロピーを用いた場合の認識結果である．表 4 と表 5 の WER の違いはわずかであった．提案法では，3,972 の全ての状態の尤度計算の 56%の尤度計算で同等の WER が得られる事から，アクティブな状態を用いた探索空間の近似は適切であると考えられる．

5. 考　　察

ここでは表 4 に対応して，提案法による動的な重み $W_k^{MI}(n)$ と従来法による重み $W_k^{MinH}(n)$，$W_k^{InvH}(n)$ の差異についての考察を行う．図 3 の一番目のグラフに入力音声波形の包絡を示した．二番目のグラフには包絡の各時刻に対応して，アクティブな状態数と，従来法での認識結果 E:と提案法による認識結果 C:を示した．三番目のグラフには，提案法により得られる三種類のストリームの統合重みを，四番目のグラフには，逆数を用いた従来法により得られる各ストリームの統合重みを示した．図示したストリーム重みの軌跡は，比較的滑らかであり，隣接フレームの尤度を用いた統合重みの平滑化は必要ないと考えられる．

図示した区間において，従来法 $W_k^{MinH}(n)$ では最も $W_k^{MI}(n)$ が大きいストリームを選択する事になり，ストリームを統合し

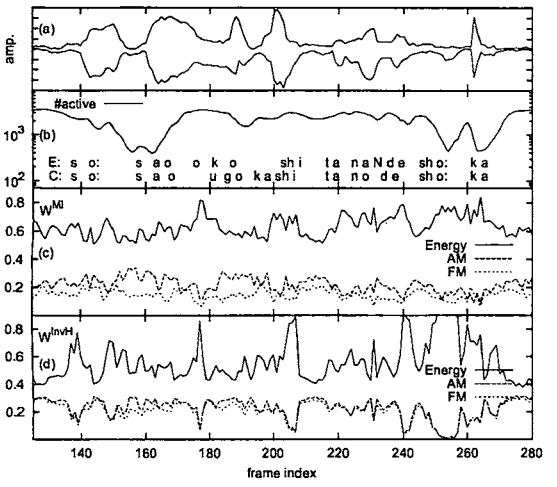

図 3 提案法による動的重み $W_k^{MI}(n)$ と，従来法による重み $W_k^{MinH}(n)$，$W_k^{InvH}(n)$ の比較．(a) 入力波形包絡．(b) アクティブノード数と認識結果．(c)$W_k^{MI}(n)$．(d)$W_k^{InvH}(n)$．

ない場合 "*Energy*" の認識結果と等しくなる．エントロピーの逆数を用いた従来法 $W_k^{InvH}(n)$ からは，"*Energy*" と同じ認識結果が得られ，E:に示した従来法での WER は 50%であった．一方，C:に示した提案法での WER は 0%であった．

提案法による動的重み $W_k^{MI}(n)$ と従来法による重み $W_k^{InvH}(n)$ の最も大きな差異の一つとして，"*AM*" と "*FM*" に与えられた重みの差異の有無が挙げられる．従来法では，"*AM*" と "*FM*" の識別能力の差異を評価できず，この両者にほぼ同じ重みが与えられたが，提案法では両者の識別能力の差異をストリーム重みの差異として与える事ができたと推測される．2.1 に述べた通り，提案法ではエントロピーを線形に写像してストリーム重みを得ている．この線形な写像方法では w_k による尤度の補正が不可欠であった．一方，従来法では，エントロピーの逆数を用いた非線形な写像によりストリーム重みが得られる．この非線形な特性により識別能力が高いストリームに過大な重みが与えられるため，従来法では w_k による尤度の補正なしに，認識精度の改善が得られた（表 3 *InvH* 挿入誤りを除いて評価した場合）が，尤度の補正を行った場合には，"*AM*" と "*FM*" の識別能力の差異が過小評価されてしまったと考えられる．

6. おわりに

本稿では，ストリームを観測する事により得られる探索仮説中のアクティブノードのエントロピーの削減量を基準に，複数のストリームの尤度を動的に統合する手法を提案した．ニュース音声の認識実験では，従来の MFCC に相当するスペクトル形状を表す特徴量に加え，フレーム内の AM，FM に着目した特徴量の統合を試みた．実験結果によると，追加した AM，FM のストリームは，単独および静的な統合方法では認識精度の改善が得られないストリームであったが，提案法によって動的に統合することで，スペクトル形状を表すストリーム単独で用いた場合の認識結果と比較して，5.4%の誤認識削減率が得られた．また，従来の動的な尤度統合手法と比較して，提案法によって 4.1%の誤認識削減率が得られた．

さらに，提案法と従来法により得られる動的な重みを比較し，提案法では，AM，FM のストリームの識別能力の差を評価できる事と，提案法では，静的な重みを用いたストリームの識別能力の補正が必須である事を明らかにした．

今後，提案法によるストリームの尤度統合をより有効に機能させるため，相補的なストリームの構成法の検討を進めていく．

謝　辞

ガンマトーンフィルタ係数についてご助言いただいた和歌山大学入野俊夫先生に感謝いたします．

文　献

[1] T. Imai, A. Kobayashi, S. Sato, S. Homma, K, Onoe, and T. Kobayakawa, "Speech Recognition for Subtitling Japanese Live Broadcasts," The 18th International Congress on Acoustics, pp. I 165–168, 2004.

[2] S. Sato, K. Onoe, A. Kobayashi, S. Homma, T. Imai, T. Takagi, T. Kobayashi, "Dynamic Integration of Multiple Feature Streams for Robust Real-Time LVCSR," Interspeech, pp. 1146–1149, 2007.

[3] 佐藤，小林，尾上，本間，今井，都木，小林，"マルチストリームのエントロピーを用いた動的探索手法，" 日本音響学会春季研究発表会，pp. 55–56, 2007.

[4] R. Haeb-Umbach, H. Ney, "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition," ICASSP, pp. I-13–16, 1992.

[5] M. J. F. Gales, "Semi-Tied Covariance Matrices for Hidden Markov Models" IEEE trans. S.A.P., vol. 7, pp. 272–281, 1999.

[6] J. Hernando, "Maximum Likelihood Weighting of Dynamic Speech Features for CDHMM Speech Recogniiton," ICASSP, pp. 1267–1270, 1997.

[7] Y. L. Chow, "Maximum Mutual Information Estimation of HMM Parameters for Continuous Speech Recognition Using the N-best Algorithm," ICASSP, pp. 701–704, 1990.

[8] G. Gravier, S. Axelrod, G. Potamianos, C. Neti, "Maximum Entropy and MCE based HMM Stream Weight Estimation for Audio-Visual ASR," ICASSP, pp. 853–856, 2002.

[9] A. Garg, G. Potamianos, C. Neti, T. S. Huang, "Frame Dependent Multi-Stream Reliability Indicators for Audio-Visual Speech Recognition," ICASSP, pp. I-24–27, 2003.

[10] H. Misra, H. Bourland, V. Tyagi, "New Entropy Based Combination Rules in HMM/ANN Multi-Stream ASR," ICASSP, pp. II-741–744, 2003.

[11] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, M. Allerhand, "Complex Sounds and Auditory Images," Proc. Auditory Physiology and Perception, pp. 429–446, 1992.

[12] A. Ando, T. Imai, A. Kobayashi, H. Isono, K. Nakabayashi, "Real-Time Transcription System for Simultaneous Subtitling of Japanese Broadcast News Programs," IEEE Trans. Broadcasting, vol. 46, no. 3, pp. 189–196, 2000.

[13] D. E. Ferguson, "Fibonaccian searching," Communications of the ACM, vol. 3, no. 12, p. 648, 1960.

[14] C. P. Cox, "A Handbook of Introductory Statistical Methods," Jhon Wiley and Sons, Inc., pp. 46–50, 1987.