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

2004－SLP－54（18）
2004／12／21

雑音モデルに基づく補正音響尤度を用いた音声認識

佐藤 庄衛† 尾上 和穂† 小林 彰夫† 今井 亨†

†NHK 放送技術研究所 〒157-8510 東京都世田谷区砧 1-10-11

E-mail: †{sato.s-gu, onoe.k-ec, kobayashi.a-fs, imai.t-mq}@nhk.or.jp

あらまし 本論文では，雑音環境下の音声の認識精度向上を目的とし，探索仮説中の音響尤度を補正する方法を提案する．提案法では，急激な変化を伴う多様な非定常雑音を取り扱うために，雑音モデルと音声モデルの尤度から求めた事後確率を定義し，これを入力音声の音響尤度の信頼度とする．認識時には，信頼度が低いフレームの音響尤度の重み小さくして探索を行う．提案法を，雑音を付加したニュース音声の認識実験に適用した結果，入力音声の S/N が低い場合（0-5dB）に単語誤認識率の改善が見られた．最大の改善は S/N が 0dB の場合に得られ，誤認識単語削減率 20%が得られた．さらに，本論文では入力音声の S/N が高い場合（10dB）の認識精度を改善するため，補正法の改良を提案する．また，提案法を雑音の多い野球中継音声の認識に適用したところ，他の雑音対策手法と併用しても，野球に関わるキーワードの検出精度が改善されることを確認した．

キーワード 音声認識，非定常雑音，雑音対策，音響尤度，探索，雑音モデル

Speech Recognition Adopting Compensated Acoustic Likelihood based on Noise Models.

Shoei SATO† Kazuo ONOE† Akio KOBAYASHI† and Toru IMAI†

† NHK Science and Technical Research Laboratories, 1-10-11 Kinuta, Setagaya-ku, Tokyo, 157-8510 Japan

E-mail: †{sato.s-gu, onoe.k-ec, kobayashi.a-fs, imai.t-mq}@nhk.or.jp

Abstract To improve recognition accuracy for speech uttered in a noisy environment, this paper proposes a new compensation method for acoustic scores in the Viterbi search. In this method, to cope with wider varieties of background noise whose characteristics change rapidly, a confidence factor is obtained as a posterior probability of speech models or a likelihood ratio between speech models and noise models. This confidence factor represents the reliability of the acoustic score for the input speech. In decoding, weight of the acoustic score at a noisy frame is reduced according to the value of the confidence factor. An experiment with broadcast news transcription showed that this method reduced word errors for input speech with lower SNR values (0-5dB). The greatest reduction of word errors, by 20%, was obtained at an SNR of 0dB. This paper also proposes a modification of the compensation, which improved the recognition performance at a higher SNR of 10 dB. The proposed method is also applied to recognition of a noisy sports program. The results showed the method improved accuracy of keywords that is important for automatic meta-data extraction.

Keyword speech recognition, noisy environment, acoustic score, compensation

1. はじめに

NHK では自動字幕作成，および将来の多彩な視聴形態の実現に欠かせない番組メタデータ作成を目的として，種々の番組音声の認識精度向上を図っている．現在，ニュース番組では，スタジオアナウンサーの原稿読み上げ部分で音声認識結果を用いた字幕作成が行われている [1]．しかし，背景雑音のある現場中継では，実用化に十分な認識精度が得られていない．一方，音声認識をはじめとする各種認識技術は，生放送のスポーツ中継番組などのメタデータの自動作成に，その利用が期待されている [2]．しかし，ニュース番組同様に，背景雑音が認識精度確保の障害となっている．

このような番組において，誤認識の原因となる背景雑音は非定常的であり，その種類も多様であるという特徴を持っている．この多様な雑音を網羅した大量の学習音声を用意することは困難であることに加え，有限の混合分布で表現される統計モデルを音響モデルに用いているため，雑音の多様性を反映させた音響モデルの学習は十分とはいえない．また，MLLR[3]のような適応化技術は，非定常雑音を含む適応データを入力音声ごとに容易に入手できないため，導入は難しい．さらに，雑音の非定常性により，雑音の特性を推定することが難しいため，雑音推定を必要とするスペクトルサブトラクション[4]やオンライン適応化手法[5]も十分な効果が期待できない．

雑音対策のためにモデルや入力に変更を加えない手法として，Weighted Viterbi Decoding が提案され，Aurora2 データベースを用いた数字認識[6]とニュース音声認識[7]で，雑音重畳音声の認識精度の改善が報告されている．これらの方法は，入力音声のフレーム

ごとに求めた信頼度尺度を導入して，音響尤度を補正しながらビタビ探索を行うアルゴリズムである．この方法には3つの利点がある：1) リアルタイムに動作する．2) 入力雑音に対応した適応用音声を必要としない．3) 入力雑音の特性を推定する必要がない．[6]で提案されている手法では，minimum statistics tracking method[8]で推定した S/N を信頼度として用いているが，この S/N 推定方法は背景雑音の統計量の変動が小さいことを仮定しているため，本稿で取り扱うような非定常雑音での効果はあまり期待できない．

本稿および[7]で提案する方法は，[6]と同様に Weighted Viterbi アルゴリズムに基づくが，信頼度の算出に音声モデルと雑音モデルの２つのモデルを用いる点が異なる．音声モデルは音響モデル(HMM)の学習に用いられたものと同じ音声から学習した GMM であり，雑音モデルは幅広く収集された種々の背景雑音種別ごとに特性を学習した GMM である．この雑音モデルは雑音のみのモデルであるため，多様なモデルを比較的少量の雑音データから容易に入手できる．これら２つのモデルから求めた音声モデルの事後確率を，音声らしさの信頼度として，音響尤度の重みを動的に変えながら探索を行う．この信頼度は入力フレームごとに算出されるため，提案法による非定常雑音への対応も期待できる．さらに，多種の雑音モデルを利用することで背景雑音の多様な特徴を積極的に利用できるため，音響モデルに反映されていない雑音に起因する認識誤りの軽減が期待される．

本稿の第２章では，雑音モデルを利用した信頼度算出方法と音響尤度の補正法を説明し，第３章では，音声モデル中の無音モデルを考慮して，信頼度算出方法の改善を図る．第４章では，雑音重畳ニュース音声の認識実験，第５章ではメタデータ制作を目的とした野球中継音声の認識実験を報告する．

2. 雑音モデルを用いた信頼度算出

図１に提案法の概要を示す．事前に学習された音声モデルと雑音モデルを用いて，フレームごとの音声モデル尤度と雑音モデル尤度から信頼度を求める．この信頼度を元に探索仮説中の音響モデルの音響尤度を補正して探索を進める．

2.1. 雑音モデルと音声モデル

信頼度算出のために事前に２種の GMM セットを学習する．そのうちの１つが雑音モデルのセットである．これは，多様な雑音が含まれるように収集した雑音データから学習されたものである．ニュース音声認識であれば，交通騒音などの現場中継雑音やスタジオ内でのペーパーノイズなどの雑音を過去のニュース番組から収集したものが適当である．スポーツ中継の認識であれば，歓声や場内アナウンスなどの雑音を過去のスポーツ中継から収集したものが適当である．収集した雑音の特性を，種類ごとの収集量によらずうまく表現するために，雑音モデルの学習データはいくつかのクラスタに分類して，モデルを学習する．クラスタリングは，雑音種別が既知の場合には雑音種ごとにクラスタを作成できるが，雑音種別が未知の場合にも雑音の音響特性を元に自動的にクラスタリングする手法[9]が利用できる．雑音モデル $\lambda_N = \{\lambda_N^1, \cdots, \lambda_N^I\}$ （I は雑音

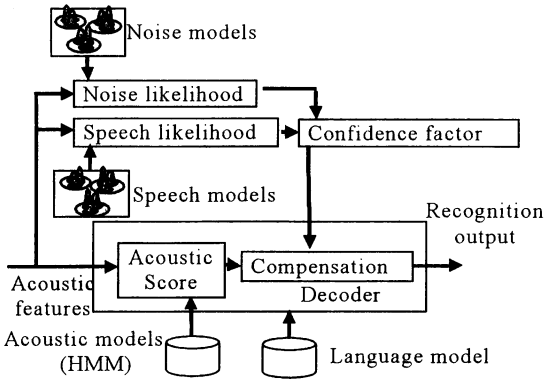

図１．提案手法の概要．

のクラスタ数）はクラスタ中の雑音データから学習された GMM である．

もう一つの GMM セットは音響モデルの学習と同じ音声で学習した音声モデルである．この GMM は，信頼度算出に必要な演算量を軽減するために，音響モデルを近似するもので，HMM に比べて小さなモデルである．雑音モデル同様，音声データもクラスタに分類することで近似精度を上げることができる．本稿では音響モデル学習時の音素ラベルを利用して，音素ごとに GMM を学習した．したがって，音声モデル $\lambda_S = \{\lambda_S^1, \cdots, \lambda_S^J\}$ （J は音声のクラスタ数）は音素 GMM となる．

2.2. 信頼度

信頼度はフレームごとに求められる．ここで，時刻 t の入力音声の特徴ベクトルを $\boldsymbol{x}_t$ とすると，音声尤度 $P(\boldsymbol{x}_t|\lambda_S)$ と雑音尤度 $P(\boldsymbol{x}_t|\lambda_N)$ は，各音声モデルの尤度 $P(\boldsymbol{x}_t|\lambda_S^j)$ と各雑音モデルの尤度 $P(\boldsymbol{x}_t|\lambda_N^i)$ をもとに，次式のように定義される．

$$P(\boldsymbol{x}_t|\lambda_S) \equiv \max_j P(\boldsymbol{x}_t|\lambda_S^j)$$
$$P(\boldsymbol{x}_t|\lambda_N) \equiv \max_i P(\boldsymbol{x}_t|\lambda_N^i) \quad (1)$$

提案法では，信頼度 γ_t に事後確率 $P(\lambda_S|\boldsymbol{x}_t)$ を用いた．

$$\gamma_t \equiv P(\lambda_S|\boldsymbol{x}_t) = \frac{P(\boldsymbol{x}_t|\lambda_S)}{P(\boldsymbol{x}_t|\lambda_S) + P(\boldsymbol{x}_t|\lambda_N)} \quad (2)$$

ここで，$P(\lambda_S) = P(\lambda_N)$ とした．

2.3. 音響尤度補正

提案法では，時刻 t に特徴ベクトル $\boldsymbol{x}_t$ が探索仮説中の各音響モデル（HMM 中の状態）から観測される確率 $P(\boldsymbol{x}_t|\lambda_{AM}^k)$ を，上述の信頼度によって補正し，探索を進める．本稿では，雑音の重畳により音響尤度が信頼できないフレーム（γ_t が小さいフレーム）では，音響尤度のダイナミックレンジが小さくなるように補正を行うため，$P(\boldsymbol{x}_t|\lambda_{AM}^k)$ を $P(\boldsymbol{x}_t|\lambda_{AM}^k)^{\gamma_t}$ で置き換えた．ここで，

γ_tは事後確率であるため$0 \le \gamma_t \le 1$である．したがって，探索仮説中の各ノードのスコア算出式は，ビタビ探索式の観測確率に信頼度で重みをつけた次式となる．

$$\alpha_t^k = \max_l \alpha_{t-1}^l \cdot a_{lk} \cdot P\left(x_t \middle| \lambda_{AM}^k\right)^{\gamma_t} \qquad (3)$$

ここで，α_t^kは時刻tの状態λ_{AM}^kでの累積音響尤度であり，a_{lk}は状態$l \to k$の遷移確率である．

この補正により，信頼度が低いフレーム$\gamma_t \to 0$では等価的に言語重みを重くして探索を行うことになる．さらに，探索の広さをビーム幅で規定した場合には，$\gamma_t \to 1$であるフレームより$\gamma_t \to 0$であるフレームの方が探索が広くなる．

3. 無音モデルに対する考察

3.1. γの計算例

提案手法により求めた信頼度を調べたところ，雑音重畳音声の非音声区間（無音区間）において信頼度が著しく低下する現象が見られた．これは音声が無い区間では，音声モデルよりも雑音モデルの尤度が高くなりやすいためであり，認識全体で無音 HMM の尤度が信頼されない傾向にある．

図２は雑音重畳音声の各フレームでの対数パワーの例である．この例は，雑音環境下の音声を模擬するため，雑音の無いニュース音声に現場中継雑音を付加して作成した音声を入力としている．図中の実線は S/N が 10dB になるように雑音を重畳した音声の対数パワーであり，破線は雑音を重畳する前の音声の対数パワーである．

図２に示した雑音重畳音声をもとに，提案法で求めた信頼度γ_tを図３に示す．図より，図２の破線で示した雑音重畳前の音声の対数パワーが小さい区間で，信頼度が小さな値になっていることがわかる．この区間は，雑音を重畳する前は無音であり，雑音の重畳によって S/N が低くなっている．

本稿では，雑音の重畳に起因して，対立仮説の音響尤度が正解仮説を上回ったために起こる誤認識の軽減を目的として，信頼度γ_tを設計した．しかし，認識すべき音声がない区間では S/N が低くなるため，無条件にγ_tが小さくなる．このような非音声区間でも，無音 HMM の音響尤度が対立仮説を大きく上回る場合に信頼度を引き上げることで，無音 HMM の尤度の過小評価を軽減できる．

3.2. γの修正

そこで本節では，無音 HMM 尤度の過小評価を軽減するため，γ_tの修正を考える．まず，λ_Sを 1)無音モデル$\lambda_S^{sil} \in \lambda_S$と 2)任意の音素のモデル$\lambda_S^{ph} \in \lambda_S \left(\lambda_S^{ph} \neq \lambda_S^{sil}\right)$の２つに分割する．ここで，$P\left(x_t \middle| \lambda_S^{sil}\right) >> P\left(x_t \middle| \lambda_S^{ph}\right)$であるフレームは，たとえ$\gamma_t$が小さな値であっても比較的信頼でき，無音区間として探索できると考え，提案法の信頼度を次式のように修正する．

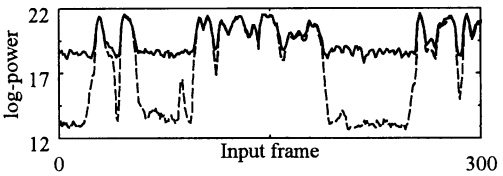

図２．入力音声の対数パワー例．実線：雑音重畳音声，破線：雑音重畳前の音声．

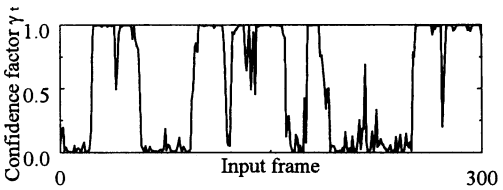

図３．雑音重畳音声から求めた信頼度．

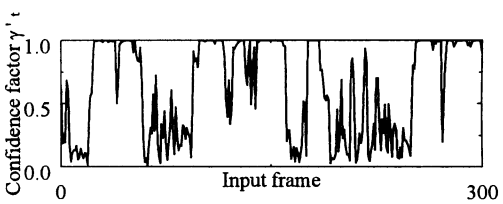

図４．雑音重畳音声から求めた修正信頼度．

$$\gamma_t' \equiv \eta_t + (1-\eta_t)\gamma_t \qquad (4)$$

修正された信頼度γ_t'は 1.0 とγ_tの重みつき和として表され，事後確率$P\left(\lambda_S^{sil} \middle| x_t\right)$を重み$\eta_t$として用いた形になっている．

$$\eta_t \equiv P\left(\lambda_S^{sil} \middle| x_t\right) = \frac{P\left(x_t \middle| \lambda_S^{sil}\right)}{P\left(x_t \middle| \lambda_S^{sil}\right) + P\left(x_t \middle| \lambda_S^{ph}\right)} \qquad (5)$$

図２に示した雑音重畳音声から求めたγ_t'を図４に示す．図３と比較すると，非音声区間でγ_t'はγ_tよりも大きな値となり，音声区間での値に大きな変化が無いことがわかる．

4. ニュース音声認識実験

4.1. 実験条件

提案法を用いて，雑音重畳ニュース音声の認識実験を行った．実験に使用したデータは全て NHK のニュース番組から収集されたデータである．評価用の音声は，雑音のないスタジオでの音声(E_S)に，ニュース番組から収集された雑音(E_N)を重畳して，S/N の異なる音声を作成したものである．評価音声および各モデル

表 1. 実験に使用したデータの詳細

	評価データ		学習データ		
	E_S	E_N	λ_{AM}	λ_S	λ_N
収集条件	背景雑音の無い音声	雑音	音声 背景雑音(あり+なし)		雑音
量	274 文 9,368 単語	786 セグメント 53 分	71,294 文 140 時間		5,276 セグメント 5 時間

の学習データの詳細を表 1 にまとめる．評価データと学習データとの間に重複はない．

音響特徴量は，MFCC12 次元および対数パワーと 1 次および 2 次の回帰係数の計 39 次元である．分析窓幅は 25msec. シフト幅は 10msec. である．この分析パラメータは音響モデル，音声モデル，雑音モデルで共通に用いた．さらに，音響モデル λ_{AM} の学習には RASTA[10]フィルタを適用し，乗算性の歪による影響を軽減した．音響モデルは，混合ガウス分布による 4,000 状態の状態共有トライフォン HMM を用い，各トライフォンのトポロジーは 5 状態 3 ループとした．また，本実験では HMM の各状態と GMM のガウス分布の混合数を 32 に統一した．

音声 GMM(λ_S)にはトライフォン HMM と同じデータから学習したモノフォン HMM の 123 状態(3 状態×無音を含む日本語 41 音素) を用いた．雑音 GMM(λ_N) はセグメントを単位として，音響的特徴をもとに自動的にクラスタリングを行った．その方法は，事前クラスタ数を設定した後，各セグメントにランダムにクラスタを割り当てた初期状態から，クラスタ内のセグメントからの GMM パラメータの学習と，GMM の尤度を基準としたクラスタリングを，混合数を増やしながら繰り返すものである．

言語モデルは，ニュース番組ごとに適応化した時期依存言語モデル[11]を用いた．この言語モデルのトライグラムテストセットパープレキシティーは 6.7．未知語率は 0.17%であった．また，提案法では，音響尤度の補正により，音響尤度ダイナミックレンジが小さくなるため，言語重みの最適値が実験条件によって異なることが予想される．そのため，言語重みは予備実験による 300 ベストのスコアを用いて，自動的に最適化を行った[12]．

4.2. γ_t による認識結果

図 5 は雑音モデルのクラスタ数(I)を 12 として，音響尤度の補正に γ_t を用いた場合の認識結果である．図は単語誤認識率(WER)を棒グラフで，音響尤度補正を行わない場合(Baseline)と比較した誤認識単語削減率(Error reduction)を折れ線グラフで S/N ごとに示したものである．図 6 は提案法による認識結果を雑音クラスタ数および S/N で比較したものである．

提案法による最大の誤認識単語削減率 20%が，雑音クラスタ数 $I=12$ の時に得られることがわかる．雑音クラスタ数 $I=16$ としたときには，雑音モデル λ_N の過学習のため，誤認識単語が増えていると考えられる．

ここで S/N が 10dB の場合，他の S/N での認識結果と比較して，2 つの特徴がみられた．1)誤認識単語削減率が，clean を除く他の S/N の雑音重畳音声に比べてきわめて小さい．2)雑音モデルのクラスタ数を増やすにしたがって，誤認識単語削減率が低下する．後者は雑音モデルの精度が向上するとともに，誤認識単語数が増えていることを意味し，提案手法による悪影響が現れたものであると考えられる．

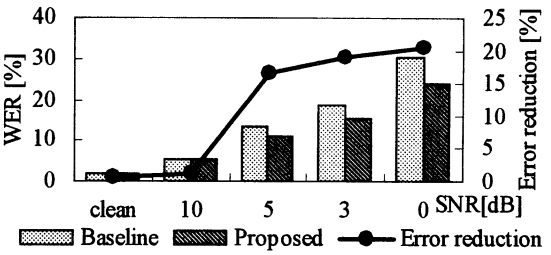

図 5. S/N ごとの単語誤認識率 ($I=12, \gamma_t$).
Baseline: 尤度補正を行わない場合．
Proposed:提案法による尤度補正．

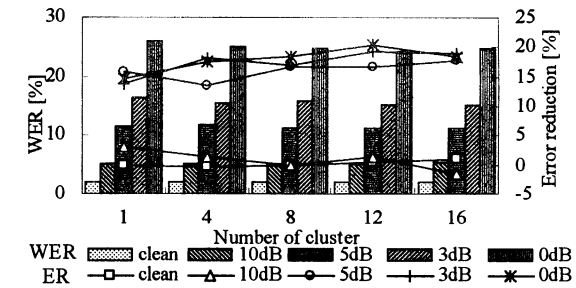

図 6. 雑音クラスタ数(I)ごとの単語誤認識率 (γ_t).

4.3. γ'_t による認識結果

図 7 は尤度補正法 γ_t と γ'_t による認識結果を S/N ごとに比較したものである．図は前節で最良の結果が得られた雑音モデル $\lambda_N(I=12)$ を用いた場合の単語誤認識率と誤認識単語削減率を，それぞれ棒グラフと折れ線グラフで示してある．γ'_t の導入によって，S/N が 10dB で誤認識単語の削減が見られる．しかし，10dB 以外の S/N では γ_t を用いた方が性能がよい．次に，S/N を 10dB に固定し，雑音モデルのクラスタ数による認識結果の比較を図 8 に示す．γ_t ではクラスタ数の増加にしたが

って誤認識単語が増加したが，γ_t'では雑音クラスタ数の増加に伴って誤認識単語が削減されることがわかる．

現状では，γ_tおよびγ_t'のどちらを採用するかは，利用するアプリケーションによって選択するのがよいと思われる．現場リポートを含む実際のニュース音声認識では10dB程度のS/Nが重要となるため，γ_t'が有効であると考えられる．

5. スポーツ中継音声の認識実験

5.1. 実験条件

生放送のスポーツ中継番組に対するメタデータ自動作成を目的として，MLB（大リーグ中継）音声の認識実験に提案手法を適用した．MLB中継の番組音声は，歓声，場内アナウンス，場内BGMなどの背景雑音により，十分な認識精度がなかなか得られない素材の一つである．

本稿で使用した評価音声は，NHKが2003年8月10日に放送したMLB中継“ヤンキース対マリナーズ戦”の放送開始から2時間分中のアナウンサーによる実況部分765文7,891単語である．

音響モデルλ_{AM}および音声モデルλ_Sは，ニュース音声認識実験とほぼ同一のものであるが，各状態の混合数は24混合に統一した．さらに，メタデータとして有用な“打撃音”，“キャッチ音”および“歓声”用の3状態1ループのイベントHMMを追加した．これらのHMMは後述の雑音データ用いて学習されたものである．雑音モデルλ_Nには，評価試合を含まないMLB中継6試合分（10時間分）から切り出された各種背景雑音データから学習した．この背景雑音データを，上記の認識対象イベント以外の主な雑音4種類に人手で分類し，分類ごとに24混合分布のGMMを学習した．雑音の分類は次のものである．1) 場内アナウンス．2) 場内BGM．3) 7イニング目のイベント(7イニング目の攻守の入れ替わり時に，歌などのイベントがある．)．4)その他．

実験に用いた言語モデルの学習には，スポーツニュースの書き起こしをベースコーパスとして用いた．さらに，適応化用のコーパスとして，評価試合を含まないMLB中継とプロ野球中継の書き起こしのチーム名，選手名，地名を評価試合の対応する名詞に置換したものを用意し，ベースコーパスに対して5倍の重みをつけて，適応化言語モデルを作成した．上述の3種類のイベントの言語確率は，間投詞（[えー][あー]など）の言語確率を等分配してエントリを作成した．作成された言語モデルのトライグラムテストセットパープレキシティーは82.3，未知語率は2.3%であった．

5.2. γ_t, γ_t'による認識結果

表2(上段)に，尤度補正法γ_t，およびγ_t'を用いた場合の単語誤認識率(WER)と，キーワード誤認識率(KER)を示す．ここで，キーワードは，メタデータの自動作成に重要であると思われる選手名と地名(90エントリ)と，“ヒット”や“ホームラン”などの野球用語(227エントリ)を用いた．評価文中のキーワード総数は1028である．

提案法の適用により，WER，KER共にわずかながら

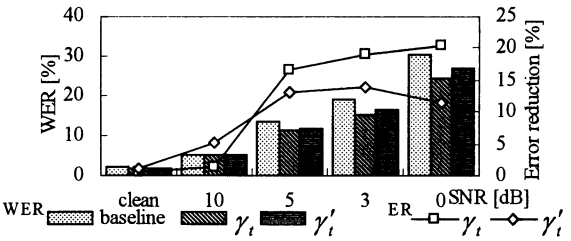

図 7. S/N ごとの尤度補正法γ_tとγ_t'の比較 ($I=12$).

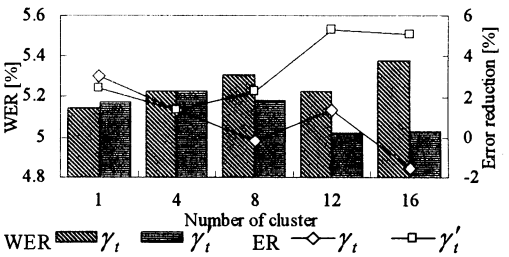

図 8. 尤度補正法γ_tとγ_t'の雑音モデル数による比較． (S/N=10 dB).

改善が見られた．本実験条件では，γ_tの方が有効であることがわかる．しかし，ニュース音声の認識実験ほどの効果は得られなかった．そこで次節では，無音HMMの尤度補正法の更なる改良を試みる．

5.3. 無音 HMM の尤度置換

MLB中継音声の場合，演出上の理由から部分的にS/Nが非常に悪くなる部分がある．また，場内アナウンスなど，その特徴が認識すべき音声の特徴に非常に近い背景雑音が存在する．このような背景雑音入力では，λ_{AM}中の無音モデルで，十分な尤度が得られない場合がある．以上のことから，雑音モデルの尤度$P(x_t|\lambda_N)$が高い部分は，入力音声が無い区間すなわち無音区間であるとして探索を行うことで，誤認識単語の削減が期待できる．実際には，雑音モデルλ_Nを音響モデルλ_{AM}中の無音モデルのバリエーションとして扱い，信頼度による重みをつけながら探索を行う．この方法によると(3)式は次式のように表現される．

$$\alpha_t^k = \max_l \alpha_{t-1}^l \cdot a_{lk} \cdot R\left(x_t|\lambda_{AM}^k\right)^{\gamma_t}$$

$$R\left(x_t|\lambda_{AM}^k\right) \equiv \begin{cases} P\left(x_t|\lambda_{AM}^k\right) & (k \neq sil) \\ \max\left\{P\left(x_t|\lambda_{AM}^k\right), P\left(x_t|\lambda_N\right)\right\} & (k = sil) \end{cases} \quad (6)$$

提案法γ_t'は，無音を考慮して(3)式のγ_tを修正したが，本節の手法は，無音モデルの尤度$P\left(x_t|\lambda_{AM}^{k=sil}\right)$を雑音モデルの尤度に置換して修正するものである．

提案する尤度置換を導入した場合の認識結果を表

表 2 MLB 中継音声の認識結果

補正法	尤度置換	WER[%]	KER[%]
×	×	47.0	45.8
γ_t	×	46.0	44.0
γ_t'	×	46.6	44.5
γ_t	○	46.1	33.7
γ_t'	○	46.5	39.3

表 3 他の雑音対策を併用した場合の MLB 中継音声の認識結果

補正法	尤度置換	WER[%]	KER[%]
×	×	32.9	29.2
γ_t	×	32.5	27.7
γ_t'	×	33.0	29.0
γ_t	○	32.4	20.5
γ_t'	○	32.8	24.4

2(下段)に示す．尤度置換により，WER の改善は見られないものの，KER を大きく改善できることがわかる．

5.4. 他の雑音対策との併用

提案手法を他の雑音対策と併用した場合の効果を調べるため，次の雑音対策を適用した場合と適用しなかった場合の比較を行った．

(1)λ_N の学習に用いた MLB 中継を適応音声として，MLLR[3]および MAP[13]による適応化を行った．

(2)乗算・加算性雑音対策として，RASTA[10]，フィルタバンクサブトラクション[4]を適用した．

表 3 は各種雑音対策手法を併用した場合について，表 2 と同様に，音響尤度補正法および無音の尤度置換の有無による単語誤認識率(WER)とキーワード誤認識率(KER)を示したものである．

各種雑音対策を併用した場合，提案手法による WER の改善は見られなかった．しかし，KER については，表 2 表同様の傾向が見られた．これは，提案手法により，今回選んだキーワードのような，長い単語の認識精度が向上し，“てにをは”などの短い単語の誤認識が増えたためであると考察される．本実験のように，メタデータ用のキーワード認識精度が必要な場合には，提案手法は各種雑音対策と併用しても効果があることが示された．

6. まとめ

本稿では，雑音環境下での音声認識の精度向上のため，雑音モデルと音声モデルの尤度比をもとに音響尤度を補正する方法を提案した．

提案手法をニュース音声認識に適応した結果，入力音声の S/N が低い場合に提案手法が有効であることが示された．実験条件の中では，S/N が 0dB の時に最も大きな誤認識単語削減率 20%が得られた．本稿では，比較的 S/N が良い場合の認識精度向上のため，修正尤度補正法を提案し，S/N が 10dB の誤認識単語削減率を，修正前の 2%から 5%に改善した．

さらに，メタデータの自動作成を目的として，MLB 中継音声の認識実験に提案法を適用し，他の雑音対策手法と併用した場合でも，キーワード認識精度を改善できることを示した．

文　献

[1] T. Imai, A. Kobayashi, S. Sato, S. Homma, K.Onoe, and T. Kobayakawa, “Speech Recognition for Subtitling Japanese Live Broadcasts,” ICA-2004, pp.I165–I168, 2004

[2] 佐野雅規，山田一郎，有安香子，住吉英樹，柴田正啓，“メタデータエディタの開発”，映メ学会年次, 7-9, 2004

[3] C. J. Leggetter, P. C. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation Continuous Density Hidden Markov Models,” Computer Speech and Language, No.9, pp.171–185, 1995

[4] K. Onoe, H. Segi, T. Kobayakawa, S. Sato, T. Imai and A. Ando, “Filter Bank Subtraction for Robust Speech Recognition,” Proc. of ICSLP, pp. 1021–1024, 2002

[5] J. C. Segura, A. de la Torre, M. C. Benitez, A. M. Peinado, “Model-based Compensation of the Additive Noise for Continuous Speech Recognition Experiments Using the AURORA II Database and Tasks,” Proc. of Eurospeech, pp.211–224, 2001

[6] X. Cui, A. Bernard, A. Alwan, “A Noise-Robst ASR Back-End Technique Based on Weighted Viterbi Recognition,” Proc. of Eurospeech, pp.2169–2172, 2003

[7] 佐藤庄衛,小早川健一,尾上和穂,本間真一,小林彰夫,今井亨，“ニュース番組における雑音モデルを利用した音響スコア補正法”，音講論，pp.31–32, September, 2003

[8] R.Martin, “Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statics,” IEEE Trans. On S.A.P., vol. 9, No. 5, pp.504–512, 2001

[9] 佐藤庄衛，世木寛之，尾上和穂，宮坂栄一，磯野春雄，今井亨，安藤彰男，“2 段階クラスタリングに基づく選択学習による音響モデル適応化”，信学論(D), Vol.J85-DII, No.2, pp.174–183,2002

[10] H. Hermansky and N. Morgan, “Rasta Processing of Speech,” IEEE Trans on A.S.S.P. vol. 6, pp.578–589, 1994

[11] A. Kobayashi, K. Onoe, T. Imai, A. Ando, “Time Dependent Language Model for Broadcast News Transcription and Its Post-correction,” Proc. of ICSLP, pp. 2435–2438. 1998

[12] R. Schwartz, S. Austin, F. Kubala, J. Makhoul, L. Nguyen, P. Placeway, G. Zayaliagkos, “New Uses for the N-Best Sentence Hypotheses within the BYBLOS Speech Recognition System,” Proc. of ICASSP, pp.1–4, 1992

[13] J. L. Guavian, C.H. Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains”, IEEE Trans. S.A.P., vol. 2, No. 2, pp. 291–298, 1994