論文 / 著書情報
Article / Book Information

論題(和文)	GMM尤度補正を用いた耐雑音音声認識
Title(English)	Noise-robust speech recognition using GMM likelihood compensation
著者(和文)	古井貞熙, 那須悠, 篠田浩一
Authors(English)	SADAOKI FURUI, Yu Nasu, Koichi Shinoda
出典(和文)	日本音響学会2011年 秋季研究発表会講演論文集, Vol. , No. , pp. 29-32
Citation(English)	Proc. Acoustical Socuiety of Japan Autumn Meeting, Vol. , No. , pp. 29-32
発行日 / Pub. date	2011, 9

GMM 尤度補正を用いた耐雑音音声認識*

☆那須悠，篠田浩一，古井貞熙（東工大）

1　はじめに

今日の統計的枠組みに基づく音声認識の技術では，理想的な環境における認識精度が高い水準に達する一方で，未解決の課題も多く残されており，雑音環境下における精度の低下がその一つである．認識精度低下の要因は，雑音によって音響特徴量が変化し，学習された音響モデルとの不一致が生じるためであり，この問題を解決するために広く研究が行われてきた．観測信号あるいは音響特徴量から雑音の影響を低減する手法 [1] や，雑音に合わせた音響モデルを使用する手法 [2] などが数多く提案され，特定の条件下で認識精度を改善できることが報告されている．しかし，これらの手法の多くは定常雑音または既知の雑音を対象としており，非定常雑音や未知の雑音に対しては高い効果が得られないことや，想定される雑音を用いて事前に学習する必要があるなどの問題点がある．

本稿では，雑音に関する事前知識を必要としない耐雑音音声認識手法を提案する．クリーン音声で学習した音響モデルを用い，隠れマルコフモデル（HMM）の各状態で出力尤度を補正することによって雑音の影響を低減する．特にフィルタバンク特徴量を用いた場合に効果が高く，MFCC を特徴量としてマルチコンディション学習を行ったときの認識精度を大きく上回る結果が得られた．

2　ミッシングフィーチャー理論

クリーン音声で学習した音響モデルを用いて雑音重畳音声を認識する手法として，ミッシングフィーチャー理論が提案されている [3–7]．ミッシングフィーチャー理論に基づく手法では，観測された特徴ベクトルのうち，雑音によって値が変化した特徴量成分を信頼できないものとしてマスクし，信頼できる成分のみを用いて認識を行う．信頼できない成分も真の特徴量の上限など範囲を制限する目的に用いられることがある．適切なマスクを用いることで，信号対雑音比（SNR）が低い場合でも大きく認識精度を改善できることが示されている．MFCC に対しては雑音がすべての次元に影響を与えるため，フィルタバンク特徴量などスペクトル空間の音響特徴量を用いる研究が多い．

多くの手法では，音声認識の前に信頼できない特徴量のマスクが推定できることを前提とする．実際の応用では，雑音の推定が難しいことと同様に，マスクの推定も容易ではない．マスクの推定精度が低いと高い認識精度を達成できないことが多くの研究で報告されている．2 値ではなく，信頼度に応じて連続値のマスクを用いる方法も提案されている [5] が，本質的な解決にはなっていない．

本稿で提案する手法もミッシングフィーチャー理論の考え方に基づくが，明示的なマスク推定は必要とせず，信頼できる成分の選択は HMM の各状態において尤度補正の形で暗黙的に行われる．

3　ガウス混合分布

音声認識で利用される HMM の各状態の出力尤度は，ガウス混合モデル（GMM）で表される．

$$\mathrm{L}_s(\boldsymbol{o}) = \sum_{m=1}^{M_s} w_{s,m}\mathcal{N}(\boldsymbol{o};\boldsymbol{\mu}_{s,m},\boldsymbol{\Sigma}_{s,m}) \qquad (1)$$

ここで，$\boldsymbol{o}$ は n 次元の観測ベクトル，M_s は状態 s の混合数，$w_{s,m}$, $\boldsymbol{\mu}_{s,m}$, $\boldsymbol{\Sigma}_{s,m}$ はそれぞれ m 番目の混合成分の重み，平均ベクトル，共分散行列である．$\mathcal{N}$ は多次元ガウス分布の確率密度関数である．音声認識で一般的に用いられる対角共分散行列の場合，多次元ガウス分布に対する対数尤度は，次元ごとの対数尤度の和

$$\log\left(\mathcal{N}(\boldsymbol{o};\boldsymbol{\mu},\boldsymbol{\Sigma})\right) = \sum_{i=1}^{n}\mathrm{LL}(o_i;\mu_i,\sigma_i), \qquad (2)$$

$$\begin{aligned}\mathrm{LL}(o_i;\mu_i,\sigma_i) &= \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{1}{2}\frac{(o_i-\mu_i)^2}{\sigma_i^2}}\right)\\ &= -\frac{1}{2}\left(\log 2\pi\sigma_i^2 + \frac{(o_i-\mu_i)^2}{\sigma_i^2}\right)\end{aligned} \qquad (3)$$

で表される．

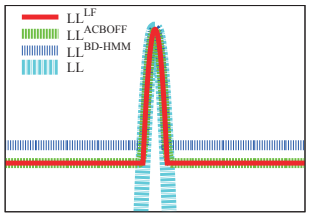
対数尤度は分布の中心と観測ベクトルとの距離の 2 乗に従って低下する．そのため，雑音によって観測ベクトルの一部の値が変化した場合，真の状態に対する対数尤度が大きく下がることがある．これが雑音環境下における音声認識精度の低下の一因になっていると考えられる．

4　従来手法

4.1　Bounded-Distance HMM

先行研究において，ミッシングフィーチャー理論の実現方法として，HMM の各状態における GMM の

*Noise-robust speech recognition using GMM likelihood compensation. by Yu Nasu, Koichi Shinoda, and Sadaoki Furui (Tokyo Institute of Technology)



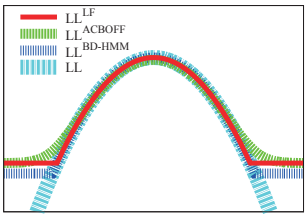


Fig. 1 Log likelihood of BD-HMM, ACBOFF and LF for two different variances.

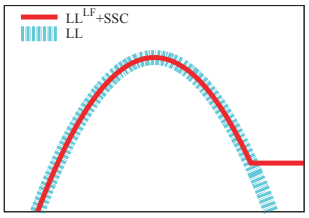


Fig. 2 Log likelihood of LF with SSC.

出力尤度または距離尺度を補正する手法が複数提案されている． Bounded-Distance HMM （BD-HMM） [6] は，クリーン音声で学習された音響モデルのガウス分布に対して，次元ごとの対数尤度計算（式 (3)）における平均からの距離に依存する項に下限を設ける手法である（図 1）．

$$\begin{aligned}&\mathrm{LL}^{\text{BD-HMM}}(o_i;\mu_i,\sigma_i)\\&=-\frac{1}{2}\log 2\pi\sigma_i^2+\max\left(-\frac{1}{2}\frac{(o_i-\mu_i)^2}{\sigma_i^2},b_i\right)\end{aligned}\tag{4}$$

ここで， b_i は次元 i における距離に依存する項の下限である．文献 [6] ではすべての次元について共通の値としているが，ここではより一般的に次元ごとに値を定めるものとする．

BD-HMM の問題点は，極端な外れ値となる特徴量に対して顕在化する．雑音による影響が大きく，いずれのガウス分布の中心 μ_i からも遠い距離を持つ特徴量に対しては， max 関数の項はすべてのガウス分布について等しい値となり，分布の分散による項の差が尤度の差となる．そのため，分散が小さいガウス分布を持つ状態の尤度が相対的に高くなる（図 1）．

4.2 Acoustic Backing-off

Acoustic Backing-off （ACBOFF） [7] は，音声の特徴量に対するガウス分布のモデルに加えて，学習データでモデル化できない特徴量に対しても一様分布を仮定して確率[1]を与える手法である（図 1）．

$$\begin{aligned}&\mathrm{LL}^{\text{ACBOFF}}(o_i;\mu_i,\sigma_i)\\&=\log\left(\alpha\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{1}{2}\frac{(o_i-\mu_i)^2}{\sigma_i^2}}+(1-\alpha)p_i\right)\end{aligned}\tag{5}$$

ここで， α はバックオフ係数， p_i は次元 i のガウス分布に加える一様分布の確率値である．

通常，音声認識における尤度の計算は対数を取った空間で行われるが， ACBOFF では尤度の空間における加算を次元ごとに行うため，対数関数の適用回数が多くなり計算コストが高くなる欠点がある．

[1] 未知のテストデータに対しては積分範囲が決められないため，厳密な確率分布ではない．

5 提案手法

5.1 Likelihood Flooring

GMM の各混合成分に対して，次元ごとの対数尤度に下限を設けて補正を行う手法， Likelihood Flooring （LF）を提案する． BD-HMM および ACBOFF と同様，クリーン音声で学習された音響モデルを用い，対数尤度の補正のみで耐雑音音声認識を行う． LF では，対数尤度を次のように計算する．

$$\begin{aligned}&\mathrm{LL}^{\text{LF}}(o_i;\mu_i,\sigma_i)\\&=\max\left(-\frac{1}{2}\left(\log 2\pi\sigma_i^2+\frac{(o_i-\mu_i)^2}{\sigma_i^2}\right),t_i\right)\end{aligned}\tag{6}$$

ここで， t_i は次元 i に対する対数尤度の下限値である．ガウス分布の平均に近い特徴量に対しては通常の対数尤度（式 (3)）と同じ値を出力する一方，平均からの距離が大きい特徴量に対しては一定の値 t_i を出力する（図 1）．

BD-HMM （式 (4)）との違いは，分布の分散による項も含めて閾値処理を行う点である．尤度の下限が分布の分散に依存する BD-HMM に対して， LF は同じ次元についてはすべてのガウス分布で尤度の下限値が等しくなるため，雑音によってどの状態のモデルからも外れる値となった特徴量に対しては状態間の尤度差がなくなり，認識への影響を抑えられる．雑音の影響が小さい次元に対しては真の状態の尤度が高く，その他の状態の尤度が低くなるため，そのような次元の特徴量が認識に使われる．

ACBOFF と比較すると， ACBOFF では式 (5) のパラメータを $\alpha \approx 1$ とするため， $\log(1-\alpha)p_i = t_i$ とおくと LF （式 (6)）とほぼ同じ値が得られる．ACBOFF の対数尤度は微分可能な関数であり， LF の境界付近のみで大きな差が生じる（図 1）．このことから， LF は ACBOFF の計算コストを削減し簡略化したものであると考えることも可能である．なお，文献 [7] では p_i の決め方は述べられておらず，パラメータとしては α についてのみ議論されている．しかし， α を固定して p_i を変化させてもほぼ同じ結果が得られ， LF と同様の効果が期待できる．

5.2 Single Side Compensation

ミッシングフィーチャー理論に基づく音声認識では，雑音環境下においても一部の特徴量が保存され，信頼できる値を持っていることが前提となる．音声認識の特徴量として広く用いられている Mel-Frequency Cepstral Coefficients（MFCC）は，メル周波数軸上で等間隔の三角窓を通した各帯域のパワーの対数であるフィルタバンク特徴量（FBANK）に対して離散コサイン変換を行ったものである．特定の帯域に局所的な雑音が音声に重畳した場合でも， MFCC では離散コサイン変換を行うために，その影響がすべての次元に拡散し，信頼できる特徴量が残るとは限らない．

一方， FBANK では帯域性雑音の影響がその周波数に対応する次元のみに限定されるため，他の次元では音声の特徴量が保存される．また対数を取るため，広帯域の雑音であっても音声のパワーが大きい次元ではクリーン音声にほぼ等しい特徴量が得られる．

FBANK のもう一つの利点として，加法性雑音による特徴量への影響がほぼ非負になるという性質がある．そのため，雑音によって値が変化した特徴量も真の特徴量の上限として利用することができる．

雑音による特徴量への影響が非負である場合，分布の平均より小さい値の特徴量が真の状態の尤度を過度に低下させることはない．従って，式 (6) による補正は $o_i > \mu_i$ となる特徴量にのみ適用すればよく，過剰な補正による認識精度の低下を抑えることができる（図 2）．以下，この手法を Single Side Compensation（SSC）と呼ぶ．本稿の実験では，特徴量に FBANK を用いる場合に静的特徴量に対して SSC を行う．

従来手法では SSC のような音響特徴量の性質の利用法は考慮されていないが， SSC は LF だけでなく ACBOFF や BD-HMM にも適用可能である．

5.3 パラメータの決定

LF （式 (6)）における閾値 t_i の決め方を述べる．各次元の対数尤度（式 (3)）は，分布平均からの距離に依存する項と分散の大きさによって決まる項からなり，分散のスケールは特徴量や学習された音響モデルによって異なるため，閾値 t_i を次式のように定める．

$$t_i = c_i + \frac{1}{N}\sum_{s=1}^{N}\sum_{m=1}^{M_s} w_{s,m}\log 2\pi\sigma_{s,m,i}^2 \qquad (7)$$

N は学習された音響モデルの状態数， c_i は距離に依存する項の補正に対応する定数である． c_i はモデルの分散に依存しない値であり， LF の閾値には学習されたモデルの全状態，全混合成分で $\log 2\pi\sigma_i^2$ を平均した値をバイアスとして加えた t_i を用いる．

6 評価実験

6.1 実験条件

実験には， CHiME corpus [8] および付属の評価スクリプトを用いた．文単位のコマンド音声に，一般家庭で収録された雑音を重畳したデータで構成される．開発セットおよびテストセットの音声には SNR が −6 dB から 9 dB となるように雑音が重畳されている．タスクはコマンド音声の認識で，連続的に発声される文のうちキーワード 2 単語の正解率で評価される．チャンスレートは 7% である． CHiME corpus ではバイノーラルの音声データが提供されているが，本稿の実験では両チャネルを平均したモノラル信号を用いた．

特徴量の抽出と学習，認識には HTK [9] を用い，特徴量は MFCC および FBANK とした．フィルタバンク解析のチャネル数は 26， MFCC の次元数は 0 次を除いて 12 とし，いずれの特徴量にも対数パワーを加えて一次および二次のデルタ特徴量を付加した．MFCC には Cepstral Mean Subtraction （CMS）を行った． HTK の表記で， 39 次元の `MFCC_E_Z_D_A` および 81 次元の `FBANK_E_D_A` となる．

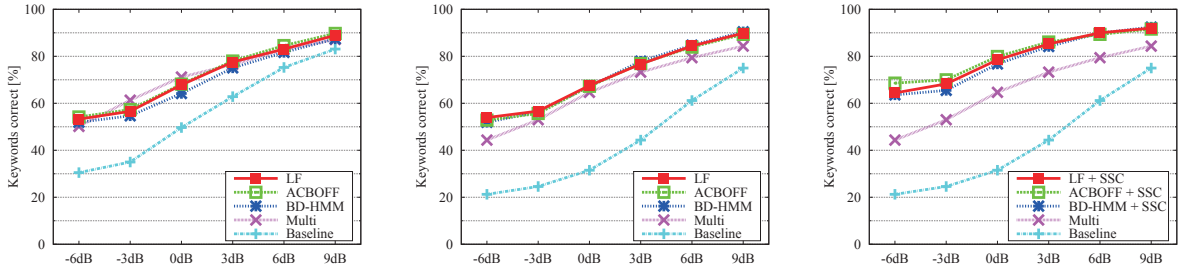
音響モデルは 34 名の話者ごとに特定話者モデルを学習した．各状態は対角共分散 7 混合の GMM である． CHiME corpus 付属のスクリプトで学習された音響モデルは，一部のガウス分布の分散が 0 になる問題があったため，分散の値は各次元において下位 1% の値でフロアリングした．その他の条件は CHiME corpus 付属のスクリプトに準じて学習を行った．

クリーン音声で学習した音響モデルを使用する手法として，提案手法（LF）のほか，通常の認識器で認識したもの（Baseline）， BD-HMM， ACBOFF を比較対象とした．フィルタバンク特徴量に対しては，静的特徴量に対して SSC を行わない場合と行う場合の両方を比較した．各手法のパラメータは， ACBOFF の α は文献 [7] で適当とされた 0.9 に固定し，提案手法と ACBOFF， BD-HMM においてそれぞれ c_i, b_i を開発セットの認識率を最大化する値に調整した．いずれの手法についても，各パラメータは 3 つの値 c_s, c_d, c_a を用いて

$$c_i = \begin{cases} c_s & \text{for static features} \\ c_d & \text{for first order delta features} \\ c_a & \text{for second order delta features} \end{cases} \qquad (8)$$

のようにした．

また，雑音を利用して学習を行う手法として，SNR が −6 dB から 9 dB となる雑音を重畳した音声でマルチコンディション学習を行って通常の認識器で認識したもの（Multi）も比較対象に加えた．



(a) `MFCC_E_Z_D_A` (b) `FBANK_E_D_A` (c) `FBANK_E_D_A`, employing SSC

Fig. 3 Results for BD-HMM, ACBOFF and LF with `MFCC_E_Z_D_A` and `FBANK_E_D_A`.

6.2 実験結果

音響特徴量として MFCC を用いたときの実験結果を図 3(a) に示す．LF，ACBOFF はいずれも Multi と同程度の改善が得られ，BD-HMM がそれらより若干低い結果となった．

FBANK を用いたときの実験結果は図 3(b) のようになり，MFCC の場合と比較して Baseline, Multi が下がっているが，LF，BD-HMM，ACBOFF では同等の精度が得られた．3 手法の間では大きな差はみられなかった．静的特徴量に対して SSC を適用すると図 3(c) に示す結果が得られ．LF，BD-HMM，ACBOFF の認識精度が大きく改善した．

6.3 考察

特徴量として FBANK を用いると，MFCC を使用した場合より Baseline, Multi の認識精度が低下した．MFCC で CMS を行っていることに加えて，FBANK では雑音の影響が少ない次元に集中するために，各次元の特徴量の値の変化に対して 2 乗で低下する対数尤度が大きく変化するためであると考えられる．LF，BD-HMM，ACBOFF によって MFCC の場合と同程度の認識精度が得られており，信頼できる特徴量は多く残っていると考えられる．

FBANK の静的特徴量に SSC を行った場合の効果はいずれの手法でも大きく，雑音に対する特徴量の性質を利用することによって認識精度が向上することが示された．ACBOFF における確率分布の解釈では，雑音重畳音声の分布として，クリーン音声の分布平均以上の値を取る一様分布を仮定することに相当し，より正確なモデル化になっていると考えられる．

本稿の実験では，尤度補正の 3 手法は効果が高い順に ACBOFF，LF，BD-HMM となった．BD-HMM の効果が LF より低い理由は，4.1 節で議論したように，分散の小さいガウス分布の尤度が高くなる偏りが生じたためであると考えられる．LF と ACBOFF による対数尤度は 5.1 節で示したように非常に近い関数であり，両手法の認識精度の差は計算コストとのトレードオフになっている．

7 まとめ

尤度補正を用いた耐雑音音声認識手法 Likelihood Flooring および Single Side Compensation を提案し，フィルタバンク特徴量を用いることで雑音環境下におけるコマンド音声の認識精度を大きく改善することを確認した．提案手法はクリーン音声で学習した音響モデルを用い，雑音の事前知識を必要としないが，MFCC を用いたマルチコンディション学習を上回る性能が得られた．また追加の計算コストも小さい．

本稿の実験では特定話者の音響モデルを用いたが，予備実験では不特定話者モデルを用いると特定話者の場合と比較して十分な改善が得られなかった．この問題の解決や，より効果的な特徴量の検討，大語彙音声認識への適用が今後の課題となる．また，従来手法の BD-HMM で雑音抑圧手法であるスペクトルサブトラクション [1] との組み合わせが効果的であることが示されている [6] ように，他の耐雑音手法との組み合わせによって認識精度の向上が期待できる．

参考文献

[1] S. F. Boll, IEEE Trans. ASSP, 27 (2), 113–120, 1979.
[2] M. Gales and S. Young, in Proc. Eurospeech, 837–840, 1993.
[3] M. Cooke *et al.*, n Proc. ICSLP, 1555–1558, 1994.
[4] B. Raj and R. M. Stern, IEEE Signal Processing Magazine, 22 (5), 101–116, 2005.
[5] J. Barker *et al.*, in Proc. ICSLP, 373–376, 2000.
[6] J. Vicente-Peña *et al.*, Speech Communication, 52 (2), 123–133, 2010.
[7] J. de Veth *et al.*, Speech Communication, 34 (3), 247–265, 2001.
[8] H. Christensen *et al.*, in Proc. Interspeech, 1918–1921, 2010.
[9] Hidden Markov Model Toolkit (HTK), `http://htk.eng.cam.ac.uk/`.