短時間雑音特性に基づく雑音下音声認識の性能推定の検討*

☆森下恵里，山田武志，牧野昭二，北脇信彦（筑波大）

1 はじめに

近年の音声認識技術の発展は目覚しいものの，一般に広く利用されているとは言い難い状況にある．主な理由の一つとして，雑音環境における認識性能の低下が挙げられる．一般のユーザは周囲が多少騒がしくても普通に認識されることを期待するが，実際には誤認識が頻出するために使い続ける意欲を失ってしまうことになる．

この問題を本質的に解決するためには，認識性能を十分高いレベルに引き上げる必要があるが，これを短期的に達成できる見通しはたっていない．そこで我々は，ユーザに音声認識が期待通りに動作するか否かを通知する機能を実現することにより，この問題の解決を図る．例えば携帯端末における音声入力を考える．周囲が騒々しい状況では認識性能が低下するので，アイコンなどを用いてユーザにその旨を通知する．それを見たユーザは，無駄な発話を行うことなく，タッチパネルなどの別の入力手段を選択できるようになる．結果的に誤認識は大幅に削減され，ユーザの音声認識に対する信頼や理解を得ることにつながる．

本稿では，その一実現手法として，雑音環境における認識性能を短時間雑音特性から推定し，その推定認識性能に基づいて発話抑制する手法を提案し，その有効性を検証する．

2 提案手法

2.1 短時間雑音特性を用いた認識性能推定

発話抑制を適切に行うためには，まさに今ユーザが発話を始めたとするときに期待できる認識性能を推定することが重要となる．一方，認識性能の推定のために用いることができるのは，マイクロホンを通して既に観測済みの雑音である．よって，観測済み区間の雑音を用いて未観測区間の認識性能を推定するという問題に取り組む必要がある．

従来，雑音環境における認識性能を推定する手法がいくつか提案されている [1]．しかし，上記の問題設定に照らし合わせると，従来手法はあくまで観測済み区間の認識性能を推定するものである．従来手法の枠組みで未観測区間の認識性能を推定するためには，未観測区間の雑音特性を知る必要があるが，一般的に雑音は非定常であることから極めて困難である．そこで我々は，雑音は長時間では非定常であるものの，極短い時間では定常とみなせるという仮定を置く．この仮定により，隣接した 2 つの区間における雑音特性を同等とみなすことが可能となり，従来手法の枠組みを適用できるようになる．ただし，区間長については際限なく短くできるわけではなく，認識性能を算出できる程度の長さが必要である．本稿では，まず音節長に相当する 200ms について検討する．これは，発話冒頭の数音素に対する推定認識性能を発話抑制の指標とすることに相当する．

提案手法では，観測済み区間の雑音特性と観測済み区間の認識性能の関係式（推定式）をあらかじめ実験的に求めておき，この推定式に雑音特性を代入することにより認識性能を推定する．本稿では次式の推定式を用いる．

$$y = \frac{a}{1 + e^{(b_1 x_1 + b_2 x_2 + \cdots + b_n x_n - c)}} \tag{1}$$

ここで，y は推定認識性能，x_i は雑音特性を表す特徴量の i 番目の要素である．a，b_i，c は定数であり，様々な雑音環境における雑音特性と認識性能を実験的に求め，両者の関係を最適近似することにより決定する．なお，雑音のスペクトル形状が認識性能に影響を及ぼす [2] ことから，本稿では特徴量として 24 次のメル対数フィルタバンクを用いることにする．

2.2 推定認識性能に基づく発話抑制

発話抑制の目的は，推定した認識性能に基づいて，ユーザの発話を抑制するか否かを判定することである．本稿では，シンプルな閾値により判定する手法を提案する．具体的には，推定認識性能が閾値未満の場合，ユーザの発話を抑制するように通知を行う．閾値を高くしすぎると，発話抑制を行わない場合に比べて認識性能が大幅に改善する一方，音声認識の利用頻度は激減する．逆に閾値を低く設定しすぎると，音声認識の利用頻度を維持する一方，認識性能はさほど改善しなくなる．このように，提案手法では認識性能と音声認識の利用頻度にトレードオフの関係が生じる．音声認識の利用頻度をさほど減らすことなく，認識性能を大幅に改善するような閾値を見出すことが重要である．

3 提案手法の有効性の検証

3.1 認識性能推定の評価

音声データとして，新聞記事読み上げ音声コーパスのテストセット 100 文を用いた．各音声データから前後の無音区間を除いた後，発話冒頭の 200ms の音声区間を切り出した．雑音データには，電子協騒音データベースの car1，hall1，train2，lift2 を用いた．各雑音データから 20s 毎に 400ms の雑音を 8 区間切出し，前半の 200ms を観測済み区間の雑音，残りの 200ms を未観測区間の雑音とした．これらの雑音データを，SNR が $20, 18, \cdots, 0$dB になるように全音声データに重畳した．ここで，SNR は全音声データの平均パワーと各区間の雑音データのパワーの比により定義される．雑音重畳音声データの総数は，観測済み区間，未観測区間共に，100（発話）× 4（雑音）× 8（区間）× 11（SNR）の 35200 である．以上の雑音重畳音声データに対して，最後の音素が子音であることを許容する音節制約付き連続音素認識を行った．音響モデルは 3 状態 16 混合分布のモノフォンモデル，特徴量は MFCC，Δ MFCC，Δパワーの計 25 次元である．

Fig. 1 に観測済み区間の雑音特性から未観測区間の音素正解精度を推定した結果を示す．ここで，各点

*Performance estimation of noisy speech recognition based on short-term noise characteristics, by Eri Morishita, Takeshi Yamada, Shoji Makino, Nobuhiko Kitawaki (University of Tsukuba).

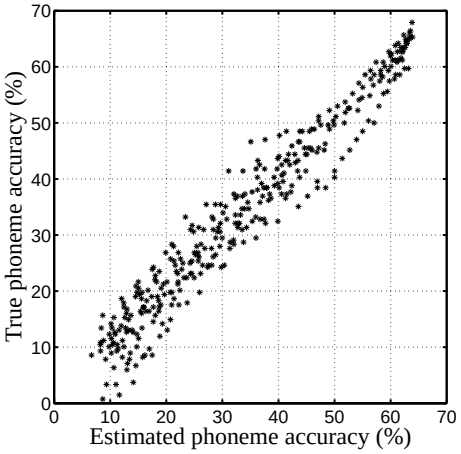

Fig. 1: Estimation result

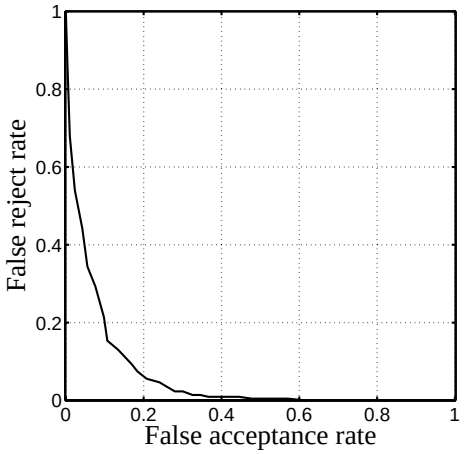

Fig. 2: ROC curve

は 352 通りの雑音重畳条件に対応する．決定係数は 0.95，RMSE は 4.01 であり，良好に推定できていることが分かる．観測済み区間の雑音特性から観測済み区間の音素正解精度を推定したときの決定係数は 0.99，RMSE は 2.16 であり，これは 2.1 節の仮定が妥当であること，及び 200ms がその仮定を満たしていることを示唆している．

3.2 発話抑制の評価

音声データとして，AURORA-2J に含まれる 150 個の一桁数字音声データを用いた．各音声データからは前後の無音区間を取り除いた．提案手法では発話冒頭の数音素に対する推定認識性能を発話抑制の指標としているが，今回発話抑制の対象とする一桁数字はそれよりもやや長い．雑音データには，電子協騒音データベースの train2 を用いた．この雑音データから 1s 毎に $200 + l_i$ms（l_i は i 番目の音声データの長さ）の雑音を 150 区間切出した．そのうち，前半の 200ms を観測済み区間の雑音とし，認識性能推定に用いた．また，残りの $l_i$ms を未観測区間の雑音とし，各音声データに重畳した．ここで，SNR は全音声データの平均パワーと雑音データ全体のパワーの比により定義され，$20, 16, \cdots, 0$dB の 6 種類である．雑音重畳音声データの総数は，1（雑音）× 150（区間）× 6（SNR）の 900 である．すなわち，受理すべき発話（正しく認識される発話）の数と抑制すべき発話（誤認識される発話）の数の和は 900 である．以上の雑音重畳音声データに対して，一桁数字の孤立単語認識を行った．その他の条件は前節と同じである．

まず，Fig. 2 に ROC 曲線を示す．ここで，受理誤り率（FAR：False Acceptance Rate）は受理すべき発話に対する正しく受理されなかった発話の割合，抑制誤り率（FRR：False Reject Rate）は抑制すべき発話に対する正しく抑制されなかった発話の割合である．FAR と FRR がおおよそ等誤り率になるのは，閾値を 20%に設定したときであった．

次に，Fig. 3 に発話抑制後の単語正解精度（すなわち受理した発話に対する正しく認識した発話の割合）と受理した発話の数を示す．ここで，横軸は閾値，左の縦軸は発話抑制後の単語正解精度，右の縦軸は受理した発話の数である．閾値が 0%のときは発話抑制なしの場合に一致し，このときの単語正解精度は 76.1%である．それに対して，閾値が 20%のとき

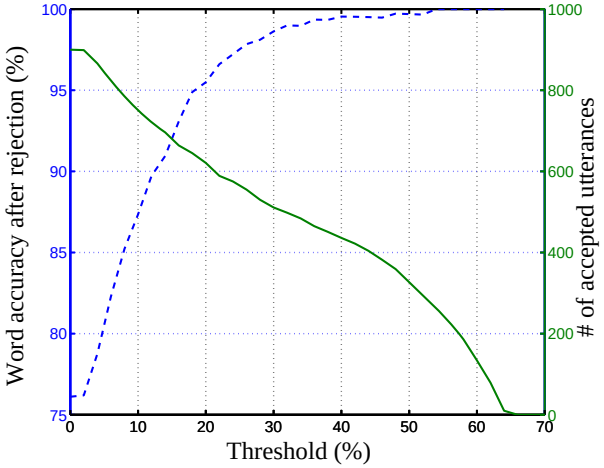

Fig. 3: Word accuracy after rejection and # of accepted utterances

の単語正解精度は 95.5%であり，発話抑制を行うことにより 19.4%の改善を得た．このときの受理した発話の数は 621 であり，これは約 3 割の発話を抑制したことに相当する．以上より，音声認識の利用頻度を著しく減らすことなく，認識性能を大きく改善できることが示された．

4 おわりに

本稿では，雑音環境における認識性能を短時間雑音特性から推定し，その推定認識性能に基づいて発話抑制する手法を提案し，その有効性を示した．今後は，最適な観測済み区間長の検討，一桁数字よりも長い発話に対する有効性の評価を行う．また，今回ユーザは発話抑制の指示に完全に従うとしているが，必ずしもそうとは限らないため，被験者実験を行う必要がある．

参考文献

[1] T. Yamada *et al.*, “Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice,” IEEE Trans. ASLP, Vol. 14, No. 6, pp. 2006–2013, 2006.

[2] 遠藤俊樹 他, “雑音の特徴分析に向けた実環境雑音データベースの構築,” 日本音響学会誌, Vol. 64, No. 1, pp. 8–15, 2008.