雑音下音声認識における必要発話音量提示機能の実装と評価*

☆後藤孝宏，山田武志，牧野昭二（筑波大）

1 はじめに

近年の音声認識技術の向上とスマートフォンの普及によって，音声認識を用いたサービスはより一般的なものとなり，様々なシーンで活用されるようになった．しかし，音声認識サービスのさらなる発展のためには，いくつかの課題を解決する必要がある．

そのような課題の一つとして，雑音環境における認識性能の低下が挙げられる．音声認識を初めて利用するユーザの多くはこのことを知らず，雑音環境下では何度も音声認識に失敗してしまう．現在のほとんどの音声認識サービスにおいては，誤認識の原因をユーザにフィードバックしていないため，ユーザはどのように対処すれば良いのか分からず，音声認識を利用する意欲を失ってしまうこともある．音声認識を使い慣れたユーザであっても，どの程度の雑音でどのくらい認識性能が低下するのかを感覚的に予測することは難しい．そのため，必要以上に大きな声で話してしまう，十分認識できる雑音環境であるにもかかわらず音声認識の利用をあきらめてしまう，といったことが起こってしまう．

この問題を解決する一つの方法は，雑音抑圧などの耐雑音手法の性能を改善することにより，認識性能を十分高いレベルに引き上げることである．しかし，耐雑音手法の有効性は対象とする雑音環境に大きく左右されることが知られている．一方，我々はこの問題をユーザインタフェース上の問題と捉え，認識できる雑音環境であるか否かをユーザにリアルタイムに通知する手法を提案した [1]．この手法では，現在の雑音環境において得られるであろう認識率を予測し，周囲が騒々しく十分な認識性能が期待できない（予測認識率が低い）場合は認識不可，静かな環境で十分な認識性能が期待できる（予測認識率が高い）場合は認識可であることをユーザに通知する．この通知に従って音声認識を利用することにより誤認識が減る，すなわちユーザが体感する認識率が向上することを示した．さらに，周囲が多少騒々しくても発話音量（入力 SNR）を大きくすれば認識できることがあるという事実に着目し，予測認識率に応じて適切な発話音量を判断して通知するように改良した手法を提案し，上記の手法よりも認識可とする雑音環境の範囲を拡大できることを確認した [2]．しかし，この手法には，適切な発話音量を判断するために 2 つの閾値を動的に設定する必要がある，認識率を予測する際に発話音量の個人差を考慮していないといった課題がある．

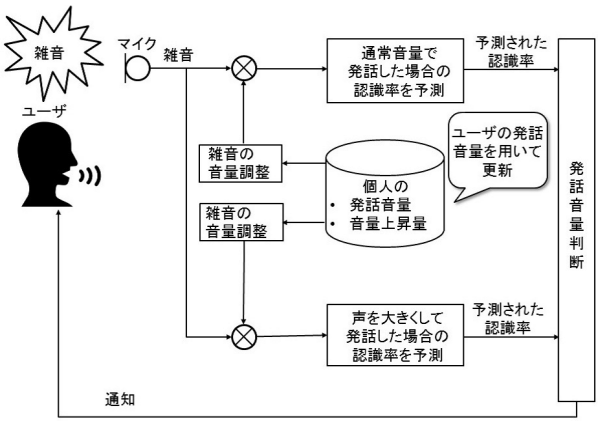


Fig. 1: Process flow of the proposed method.

そこで本稿では，静的な 1 つの閾値のみを用いて適切な発話音量を判断し，また発話音量の個人差を考慮して認識率を予測するように改良した手法を提案し，実験によりその有効性を検証する．また，提案手法をタブレット端末上に実装し，実環境における動作確認を行う．

2 提案手法

2.1 処理の流れ

提案手法の処理の流れを Fig.1 に示す．まず，直前に収録した一定長の 雑音信号を用いて，現時点で得られるであろう認識率を予測する．この予測認識率は，通常の音量で発話した場合と大きな音量で発話した場合のそれぞれに対して求める．その際，発話音量の個人差を反映させるために雑音信号の音量を調節する，次に，これら 2 種類の予測認識率から適切な発話音量を判断し，ユーザに通知する．ここで，発話音量は，音量通常，音量大，認識不可の 3 段階に分けられている．これはユーザに通知する内容に対応している．

以下では，認識率予測，発話音量判断，ユーザ音量更新について詳しく述べる．

2.2 認識率予測

認識率予測では，直前に収録した一定長の雑音信号から特徴量を抽出し，現時点において得られるであろう認識率を SVR（Support Vector Regression）[4]

*Implementation and evaluation of notification of utterance volume required in noisy speech recognition, by Takahiro Gotou, Takeshi Yamada, Shoji Makino (University of Tsukuba).

により予測する．予測する認識率は，通常の音量で発話した場合と大きな音量で発話した場合の 2 種類である．大きな音量で発話した場合の認識率予測については，通常の音量に対する上昇量に応じて，予測に用いる雑音信号の振幅を小さく調整することにより行う．さらに，認識率予測は SVR の学習に用いた話者の平均的な発話音量を基準とするため，この基準音量とユーザの通常音量との音量差に応じて雑音信号の振幅を調整する．なお，ユーザの音量については，ユーザの過去の数発話を用いて算出し，発話毎に更新する．

本稿では，雑音信号からメルフィルタバンク出力 24 次元と対数パワー 1 次元の計 25 次元を算出する．ここで，雑音長は 2 秒，分析フレーム長は 25ms，フレーム周期は 10ms である．計 197 個のフレームから特徴量を抽出し，その平均値と分散値からなる計 50 次元を認識率予測のための特徴量とする．認識率予測に用いる SVR の学習のための教師データは，種々の雑音条件の下で音声認識を行うことにより得られた単語認識率である．特徴量データは，各雑音条件に対応する雑音信号から求めた特徴量である．

2.3 発話音量判断

発話音量判断では，2 種類の予測認識率と静的な 1 つの閾値を用いて，適切な発話音量を判断する．ここで，発話音量は，音量通常，音量大，認識不可の 3 段階に分けられている．これはユーザに通知する内容に対応している，具体的な判断方法は以下の通りである．

- 通常の音量で発話した場合の予測認識率と大きな音量で発話した場合の予測認識率が共に閾値より大きいとき，通常音量と判断する（通常音量で認識できる）．
- 通常の音量で発話した場合の予測認識率と大きな音量で発話した場合の予測認識率が共に閾値より小さいとき，認識不可と判断する（音量を大きくしても認識できない）．
- 通常の音量で発話した場合の予測認識率が閾値よりも小さく，大きな音量で発話した場合の予測認識率が閾値より大きいとき，音量大と判断する（音量を大きくすれば認識できる）．

2.4 ユーザ音量更新

ユーザ音量更新では，ユーザの音量を過去の数発話から算出し，また発話毎に更新する．まず，通常音量という通知を行った場合と音量大という通知を行った場合を区別し，それぞれについて音量を算出する．次に，現時点の音量と過去の数発話の音量の重み付き平均として，それぞれの音量を更新していく．これにより，2.2 節で述べた基準音量との差を補正することが可能となる．また，ユーザの音量はシーンによって大きく変化すると考えられるが，それに追随することが可能となる．

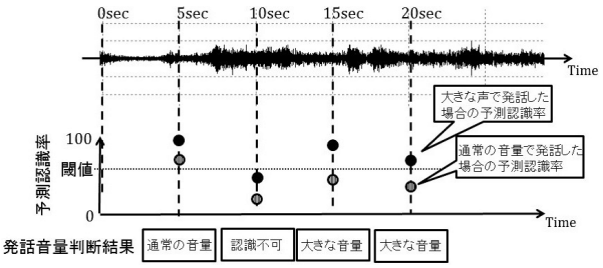


Fig. 2: Overview of the simulation experiment.

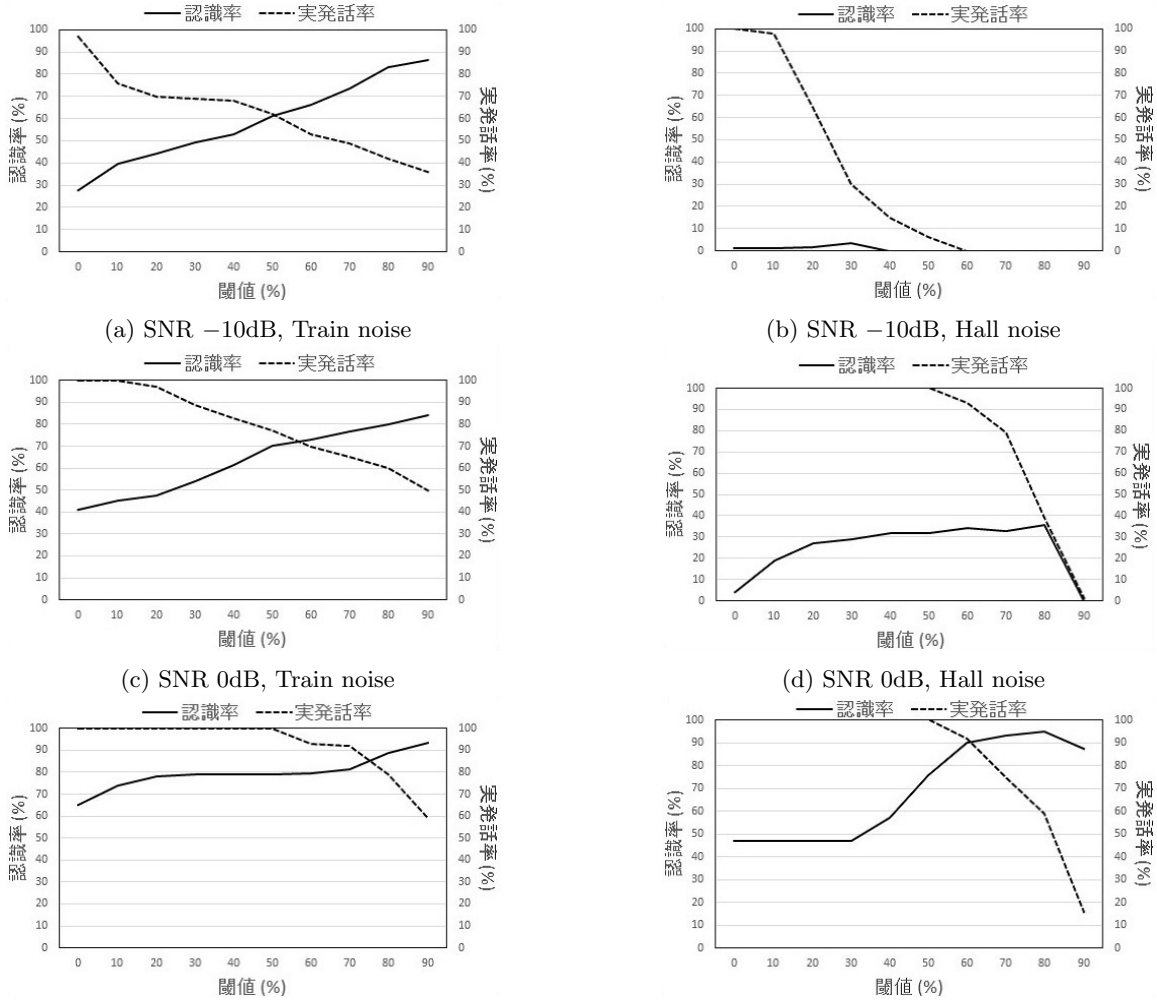
3 提案手法の有効性の評価

3.1 実験条件

提案手法の有効性を確認するためシミュレーション実験を行った．実験の概要を Fig. 2 に示す．長時間の雑音に対して 5 秒間隔で 100 回の発話機会を設定し，各発話機会において認識性能を予測し，3 段階の発話音量の通知を行う．ユーザは必ずその通知に従うものとし，通常の音量で発話する，大きな音量で発話する，発話しない，のいずれかの行動をとる．ただし，個々のユーザの通常発話音量は予め分かっているものとして，認識率予測に用いる雑音信号の音量調整を行った．また，閾値は 0%から 90%まで 10%刻みで変化させた（100 回の発話機会の全てにおいて共通の閾値を設定）．

次に，実験に用いた音声と雑音について述べる．音声データには，東北大松下単語音声データベース [5] から男女各 2 名の音韻バランス 400 単語（1 名あたり 100 単語）を用いた．話者毎に上記のシミュレーションを実施し，100 回の発話機会のそれぞれに対して重複しないように 1 単語の音声データを割り当てた．雑音データには，電子協騒音データベース [6] から走行自動車内，展示会場（ブース内），列車（在来線），ホール（百貨店）の 4 種類を用いた．これらの雑音と音声を重畳することにより，雑音重畳音声データを作成した．重畳の際の SNR は 10dB，0dB，−10dB であり，大きな音量で発話する場合にはそれぞれ 8dB 高くする．ここで，SNR は，100 単語の音声データの平均パワーと長時間の雑音データのパワーの比によって定義される．よって，各発話機会における SNR は雑音の時間特性によって異なる．

音声認識は，辞書サイズ 106 の孤立単語認識である．音響モデルは状態数 3，混合分布数 16 のモノフォン



(a) SNR −10dB, Train noise

(b) SNR −10dB, Hall noise

(c) SNR 0dB, Train noise

(d) SNR 0dB, Hall noise

(e) SNR 10dB, Train noise

(f) SNR 10dB, Hall noise

Fig. 3: Result of the simulation experiment.

HMM である．この HMM を研究用連続音声データベース（ASJ-JIPDEC）[7] と新聞記事読み上げ音声コーパス（ASJ-JNAS）[8] を用いて学習した．認識率予測のための SVR は，上記の音声データと雑音データを用いて学習した．SVR のカーネルは RBF カーネル，コストパラメータは 750 とした．なお，SVR ツールとして LIBSVM[9] の epsilon-SVR を利用した．

3.2 実験結果

ある 1 名の話者における閾値を変化させたときの認識率と実発話率の関係を Fig. 3 に示す．ここで，雑音は列車（在来線）とホール（百貨店）である．なお，走行自動車内と展示会場（ブース内）に対する実験結果は，それぞれ列車（在来線），ホール（百貨店）と同じ傾向が見られた．また，実験結果の傾向は話者間で同じであった．

まず，列車（在来線）に対する実験結果について述べる．閾値が 0%のとき，全ての発話機会において通常音量で発話することになり，SNR が 0dB の場合の認識率は約 40%である．閾値を大きくしていくと，認識率は徐々に改善していき，また実発話率は低下していくことが分かる．これは，予測認識率が低いときに認識不可，あるいは大きな音量で発話，と通知することにより，誤認識される発話を削減したことによる．他の SNR の場合も同様の結果が得られていることが確認できる．

次に，ホール（百貨店）に対する実験結果について述べる．SNR が −10dB の場合，閾値を 0%から大きくしていくと実発話率が急激に低くなっていることが見て取れる．これは，ホール（百貨店）が認識を困難とする雑音であるために予測認識率が総じて低くなり，ほとんどの発話機会において認識不可と通知するからであり，提案手法は正しい動作をしていると言える．また，SNR が 0dB の場合，閾値を大きくしていくと認識率が緩やかに改善していくことが分かる．しかし，元の認識率が極めて低いために，改善効果は限定的となっている．最後に，SNR が 10dB の場合は，列車（在来線）と同様に高い効果が得られている

Fig. 4: Notification icon[2]

ことが確認できる．以上のことから，提案手法は有効であると言える．

4　提案手法の実装

携帯型の端末を様々な雑音環境に持ち運んで音声認識を利用することを想定し，提案手法をタブレット端末（Google 社の Nexus 7）上に実装した．開発環境には Android Studio，プログラム言語には java を用いており，2 章で述べた仕様に従って実装した．ユーザへの 3 段階の発話音量の通知については，Fig. 4 に示すようなアイコンと文字を併用して行うようにした．ここで，アイコンを更新する時間間隔は，通知の分かり易さにかかわる重要なパラメタである．更新間隔が短すぎる場合には通知内容が頻繁に切り替わるため，ユーザは判断し難くなる．逆に更新間隔が長すぎる場合には雑音環境の変化に追従することができず，通知内容にミスマッチが生じることになる．そこで，本実装では，アイコンを更新する時間間隔を任意に設定できるようにした．

提案手法はあくまで音声入力の補助機能であるため，提案手法の有効性を確認するために，音声認識を行って結果を表示するというシンプルなアプリケーションを作成した．ここで，音声認識には Google 音声認識を利用している．

上述したタブレット端末を用いて提案手法の有効性を検証した．実験は防音室で行った．雑音環境を疑似的に再現するため，スピーカーを壁の角に向けて設置し，3.1 節で述べた列車（在来線）とホール（百貨店）の雑音を再生した．雑音の音量は，タブレット端末のマイクの位置での音声パワーが 55dB になることを想定し，SNR が 0dB，10dB となるように調節した．被験者は，机の上に置いたタブレット端末から少し離れた位置で単語リストにある 10 個の単語をそれぞれ発話する．その際，全ての単語が正しく認識されるまで発話を繰り返し行い，発話が終了したときの認識率と発話回数を算出する．なお，単語リストは，3.1 節の 100 個の単語の中からランダムに選択した．

提案手法を用いる場合と用いない場合の認識率と総発話回数を Table 1 に示す．表より，シミュレーション実験と同様に，提案手法を用いることによって認識率が総じて改善していることが確認できる．

Table 1: Result of the actual experiment (Recognition rate / Total number of Utterances).

	提案手法を用いない場合	提案手法を用いる場合
列車（10dB）	83% / 12 回	100% / 10 回
列車（0dB）	83% / 12 回	77% / 13 回
ホール（10dB）	71% / 14 回	91% / 11 回
ホール（0dB）	63% / 15 回	100% / 10 回

5　おわりに

本稿では，静的な 1 つの閾値のみを用いて適切な発話音量を判断し，また発話音量の個人差を考慮して認識率を予測するように改良した手法を提案し，シミュレーション実験により有効性を示した．また，提案手法をタブレット端末上に実装し，シミュレーション実験と同様に効果があることを確認した．

参考文献

[1] 森下恵里 他，“短時間雑音特性に基づく雑音環境下音声認識の性能予測の検討,”音講論，pp. 151–152，Mar. 2012.

[2] 青木智充 他，“認識性能予測に基づく雑音環境下音声認識のユーザビリティ改善の検討,”音講論，pp. 133–136，Mar. 2015.

[3] 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, Nov. 2006.

[4] AJ. Smola, B. Scholkopf, “A tutorial on support vector regression,” Statistics and computing, Vol. 14, No. 3 ,pp. 199–222, 2004.

[5] http://research.nii.ac.jp/src/TMW.html.

[6] http://research.nii.ac.jp/src/JEIDA-NOISE.html.

[7] 小林哲則 他，“日本音響学会研究用連続音声データベース,” 日本音響学会誌，Vol. 48，No. 12，pp. 888–893，1992.

[8] K. Itou *et al.*, “JNAS Japanese speech corpus for large vocabulary continuous speech recognition research,” J. ASJ (E), Vol. 20, No.3, pp. 199–206, May 1999.

[9] C. Chih-Chung *et al.*, LIBSVM: A library for support vector machines ACM Trans. Intelligent Systems and Technology, Vol. 2, No. 3, pp. 1–27, 2011.