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*

☆青木智充, 山田武志, 宮部滋樹, 牧野昭二, 北脇信彦（筑波大）

1　はじめに

一般に，雑音環境下では音声認識の性能が低下する．しかし，音声認識を初めて利用するユーザの多くはこの事実を知らず，雑音環境下では何度も音声認識に失敗してしまう．その結果，音声認識を利用する意欲を失ってしまうこともある．音声認識をある程度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かを感覚的に判断することは難しい．この問題を解決する一つの方法は，耐雑音手法の性能を更に向上させることである．しかし，耐雑音手法は対象とする雑音環境により効果が左右されることが知られている．

耐雑音手法とは異なるアプローチとして，雑音環境下での予測認識率をユーザにリアルタイムに通知することにより，誤認識（無駄な発話）を未然に防ぐ手法が提案されている [1, 2]．この手法では，周囲が騒々しく十分な認識性能が期待できない（予測認識率が閾値より小さい）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（予測認識率が閾値より大きい）場合は認識可であることをユーザに伝える．ユーザがこの指示に従うことにより，体感での認識率が向上することが示された [1, 2]．従来手法には，体感認識率を高めるために閾値を大きくすると，認識不可となる頻度が高くなってしまうという問題がある．また，最適な閾値をどのように設定するかは検討されていなかった．

そこで本稿では，予測認識率が低い場合にユーザに発話音量（入力 SNR）を大きくするように通知することにより，認識不可となる頻度を低くし，かつ体感認識率を高く保つ手法を提案する．また，ユーザの要望や雑音環境に合わせて閾値を自動的に設定する手法を提案する．最後に，実験により提案手法の有効性を示す．

2　従来手法の概要と問題点

従来手法の処理の手順を説明する．まず，常時観測している雑音の，現時刻から過去の一定長の雑音を入力とする．次に，この雑音から特徴量を抽出し，現時刻で期待できる認識率を予測する．そして，予測した認識率が予め設定した閾値を超えるかどうかにより音声認識が利用可能か否かを判断する．最後に，アイコンなどを用いてユーザへ認識可否の通知を行う．以上の処理を短時間で繰り返すことにより，ユーザは音声認識の利用可否をリアルタイムに知ることができる．ここで，ユーザが音声認識を利用しようとし，通知を見て発話するか否かを判断する機会のことを発話機会，また発話機会のうち実際に発話した割合を実発話率，実際に発話した中で認識に成功した割合を体感認識率と呼ぶことにする．

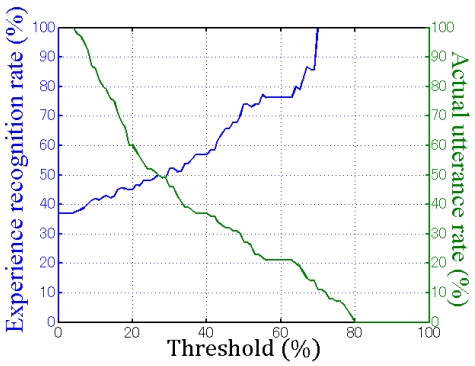

Fig. 1: Example of the relationship among threshold, actual utterance rate, and experience recognition rate in the conventional method.

従来手法を適用した場合の閾値，実発話率，体感認識率の関係の例を Fig. 1 に示す．ここで，横軸は閾値，青色の縦軸は体感認識率，緑色の縦軸は実発話率を表す．この例は，走行自動車内において 5 秒毎に 100 回の発話機会を設けたシミュレーション実験の結果である．ユーザは各発話機会で，認識可の通知時に発話を行い，認識不可の通知時には発話を行わない．

Fig. 1 から，閾値を大きくすることで体感認識率が向上することが分かる．その一方で，実発話率は低下していく．このように，実発話率と体感認識率はトレードオフの関係となっている．体感認識率が向上しても，実発話率が低くなりすぎてしまうと意味をなさない．よって，体感認識率を高く保ったまま実発話率を高めることが課題となっている．また，実発話率と体感認識率が共に高くなるような最適な閾値は雑音環境によって変わるが，どのように設定するかは検討されていなかった．

3　提案手法のアプローチと処理の流れ

予測認識率がある程度低くても，認識可と判断することを考える．このとき，体感認識率が低くなってしまうため，ユーザに発話音量（入力 SNR）を大きくする工夫を促す．このように発話音量を大きくすることで，予測よりも高い認識率が期待できる．これを実現するために，提案手法では，従来手法における認識可否判断の代わりに，発話音量判断を導入する．ここで，発話音量判断では，予測認識率が十分高い場合には普通の音量で発話をすれば認識できると判断する．また，予測認識率がかなり低い場合は認識不可（発話音量ゼロ），それ以外の場合には大きな音量で発話する必要があると判断する．このような判断の結果をユーザに通知することにより，体感認識率を高

*Usability improvement of noisy speech recognition based on recognition performance prediction, by Tomomitsu Aoki, Takeshi Yamada, Shigeki Miyabe, Shoji Makino, Nobuhiko Kitawaki (University of Tsukuba).

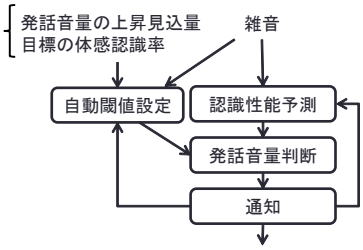

Fig. 2: Process flow of the proposed method.

く保ったまま，実発話率を高めることができると考えられる．

ここで，予備実験を行い，6 人の被験者による普通の音量の発話と大きな音量の発話を収録し，その上昇量を調べた．音量を大きくする方法として，「大きな声で話す」という指示と「マイクを口に近づけて話す」という指示を与えた．その結果，発話音量は平均 8dB 上昇し，最低でも 2dB 上昇することを確認した．この結果より，個人差があるものの，ユーザにさほど大きな負担を感じさせることなく，認識率をある程度改善できる発話音量上昇量が見込めると考えられる．

提案手法の処理の流れを Fig. 2 に示す．まず，観測した雑音から認識性能予測を行う．次に，予測認識率を基に発話音量判断を行う．最後に，発話音量判断の結果をユーザへ通知する．ユーザはそれに従って行動する．ここで，発話音量判断に用いる閾値は，ユーザの要望や雑音環境に合わせて，発話機会毎に自動的に設定する．以下，提案手法の各処理について述べる．

認識性能予測　認識性能予測では，過去のある程度の長さの雑音特徴量から現時刻における認識率を SVR (Support Vector Regression) [3] により予測する．本稿では，雑音特徴量として，メルフィルタバンク出力 24 次元と対数パワーの計 25 次元を求める．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を求めて，計 50 次元の特徴量とする．

認識率予測モデルの学習における教師データは，各発話機会で実際に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，各教師データに対応する発話機会の雑音区間から求めた特徴量である．

発話音量判断　発話音量判断では，予測認識率を基にユーザの発話音量を判断する．Fig. 3 を用いて説明する．横軸は時間，縦軸は予測認識率であり，ある雑音環境において時間とともに予測認識率が変化していることを表している．この予測認識率を閾値 A，閾値 B という 2 つの閾値により 3 クラスに分類する．ここで，予測認識率 $<$ 閾値 B のとき認識不可（発話音量ゼロ）のクラス，また 閾値 $B \leq$ 予測認識率 $<$ 閾値 A のとき大きな音量で発話するクラス，閾値 $A \leq$ 予測認識率 のとき，普通の音量で発話するクラスだと判断する．

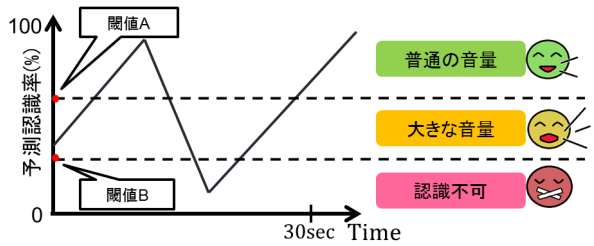

Fig. 3: Example of utterance volume classification in the proposed method.

自動閾値設定　自動閾値設定では，事前にユーザが設定した目標の体感認識率，ユーザの発話音量上昇見込量，および発話直前のある程度の長さの雑音の特徴量から，SVR を用いて閾値 A，閾値 B の設定を行う．ユーザの発話音量上昇見込量とは，大きな音量で発話するように通知されたとき，実際にはどのくらい音量を大きくできるかを指定するためのものである．これは，個人差のみならず，周囲に人がいるので大きな声を出せないという状況などに対応するために必要となる．本稿では，雑音特徴量として，発話直前の 5 秒の長さの雑音から求めたフレーム対数パワーの平均，分散，尖度，歪度を用いる．ここで，閾値の設定基準は，目標の体感認識率を達成し，かつその中で実発話率を最も高くするというものである．詳細は 5 章で述べる．

本稿では，閾値 A，閾値 B それぞれに対して自動閾値設定モデルを用意する．教師データには後述の最適な固定閾値を用いる．特徴量データには，目標の体感認識率，ユーザの発話音量上昇見込量，および発話直前のある程度の長さの雑音特徴量を用いる．

4　提案手法と従来手法の比較評価

4.1　実験条件

従来手法の認識可否判断と提案手法の発話音量判断を比較するため，閾値 A，閾値 B，体感認識率，および実発話率の関係を調査する．実験の概要を Fig. 4 に示す．図の横軸は時間であり，波形は長時間の雑音を表す．この雑音において 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通知を行う．ユーザはその通知に従い，普通の音量で発話する，大きな音量で発話する，発話しないのいずれかの行動を取る．閾値 A に関しては 0%から 100%まで変動させ，閾値 B に関しては 0%から閾値 A の間で変動させる．ここで，閾値 A，閾値 B は 100 回の発話機会において固定であり，自動閾値設定は用いない．

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 [4] から男女各 10 名の音韻バランス 106 単語を用意し，その中から 100 単語（1 名あたり 5 単語）を用いた．各発話機会には 1 名の 1 単語を割り当てた．また，雑音データとして，電子協騒音データベース [5] 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．この音声と雑音を重畳することにより，雑音重畳音声データを作成し

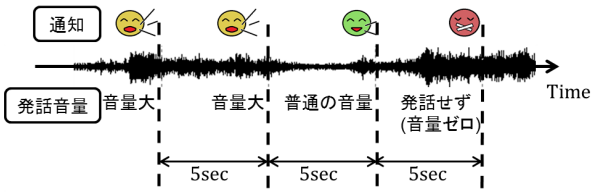

Fig. 4: Overview of the simulation experiment.

た．音声と雑音の SNR は 10dB，0dB，−10dB であり，大きな音量で発話する発話機会においては SNR を 2dB，または 8dB 高くした．ここで，SNR は全音声データの平均パワーと各雑音データの全区間の平均パワーの比により定義される．よって，雑音の時間特性に応じて発話機会毎に SNR は異なる．

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 HMM である．この HMM を研究用連続音声データベース（ASJ-JIPDEC）[6] と新聞記事読み上げ音声コーパス（ASJ-JNAS） [7] を用いて学習した．

認識率予測モデルは，上記の音声と雑音を用いて学習した．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[8] の epsilon-SVR を利用した．

4.2 実験結果

雑音が列車，SNR が −10dB，閾値 A が 80%のときの，提案手法と従来手法の体感認識率を Fig. 5 に示す．横軸が閾値 B，縦軸が体感認識率を表す．Proposed（+8dB）と Proposed（+2dB）は提案手法であり，+8dB と +2dB は発話音量の上昇量を表す．また，Conventional は従来手法を表す．図から，Proposed（+8dB）と Proposed（+2dB）は Conventional よりも，体感認識率が総じて高くなっていることが分かる．特に，Proposed（+8dB）では閾値 B を小さくしたときも体感認識率が高く保たれていることが分かる．

次に，提案手法と従来手法の実発話率を Fig. 6 に示す．横軸が閾値 B，縦軸が実発話率を表す．この曲線は Proposed と Conventional で完全に一致する．図から，閾値 B を小さくすることにより，実発話率が向上することが分かる．

以上のことから，提案手法により，体感認識率を高く保ったまま，実発話率を改善できることが確認された．

5 提案手法における自動閾値設定の評価

5.1 実験条件

発話音量判断における閾値 A，閾値 B を動的に（発話機会毎に）自動設定する場合と，最適な固定閾値を用いる場合との比較実験を行った．実験条件は基本的に 4.1 節と同じである．

ここで，最適な固定閾値とは，目標体感認識率を達成した閾値の中で，実発話率が最大となる閾値である．最適な固定閾値の例を，Fig. 7 を用いて説明する．横軸は閾値 B，縦軸は体感認識率，各曲線は閾値

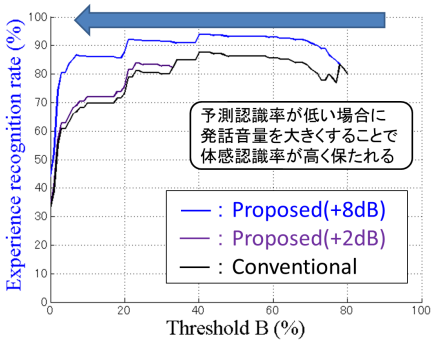

Fig. 5: Relationship between threshold B and experience recognition rate (Train noise, SNR −10dB, Threshold A 80%).

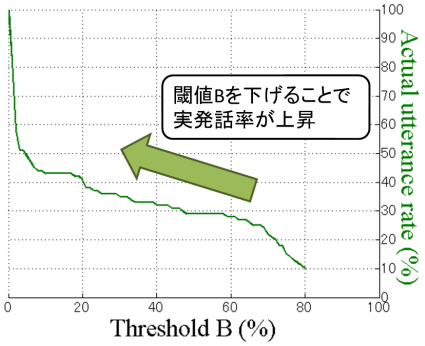

Fig. 6: Relationship between threshold B and actual utterance rate (Train noise, SNR −10dB, Threshold A 80%).

A をある値に設定したときの体感認識率の変化を表す．閾値 A は 0%, 10%, ··· , 90%まで 10%間隔で設定している．なお，閾値 A は，目標の体感認識率以上の値はとらないとしている．この制約がない場合，閾値 A は 100%となることが多く，これは常に声を大きくすることとなり，ユーザの負担を大きくしてしまう．図から，目標の体感認識率を 90%とするとき，それを達成できるのは閾値 A が 90%の場合と 80%の場合の 2 つであることが分かる．この中で実発話率が最大となるのは，閾値 A が 90%，閾値 B が 7%のときである（閾値 B が小さいほど実発話率が高い）．これが最適な固定閾値となる．

自動閾値設定モデルの学習に用いる最適な固定閾値には，4 章の実験結果を利用した．SVR のカーネルは RBF カーネル，コストパラメータは 100 とした．

5.2 実験結果

目標の体感認識率を 90%，ユーザの発話音量上昇見込量を 8dB，雑音を走行自動車内とホールとした場合の実験結果について述べる．なお，他の実験条件についても実験結果の傾向は同じである．

まず，提案手法（最適な固定閾値）の閾値 A，閾値 B を Table. 1 に示す．ここで，目標の体感認識率を達成する閾値がなかった場合，N/A としている．ホール雑音の −10dB，0dB が N/A となっているが，これは，雑音条件が厳しく，音声認識が使えるレベルにないことを示している．この表から，雑音条件によっ

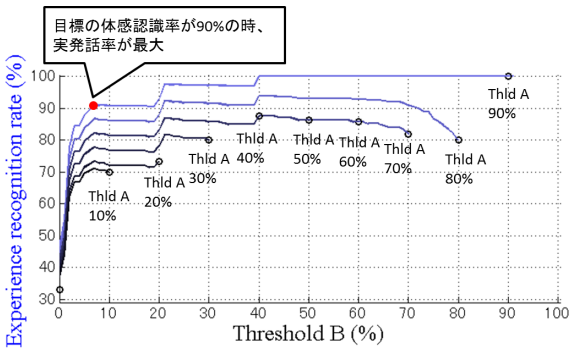

Fig. 7: Examples of optimal fixed threshold.

て最適な閾値 A，閾値 B が異なることが分かる．この閾値を手動で設定するのは困難であるため，自動閾値設定は必要不可欠である．

次に，提案手法（動的な自動閾値）と提案手法（最適な固定閾値）を用いた場合の体感認識率を Fig. 8 に示す．ここで，縦軸が体感認識率であり，横軸が雑音条件である．また，提案手法（最適な固定閾値）において目標を達成する閾値がなかった場合，および提案手法（動的な自動閾値）において目標の体感認識率が達成できなかった場合，N/A としている．図より，提案手法（動的な自動閾値）は提案手法（最適な固定閾値）と同等の体感認識率を得られることが分かる．提案手法（最適な固定閾値）の体感認識率は理想値であるが，提案手法（動的な自動閾値）はそれに匹敵する性能を達成した．

最後に，提案手法（動的な自動閾値）と提案手法（最適な固定閾値）を用いた場合の実発話率を Fig. 9 に示す．ここで，縦軸が実発話率であり，横軸が雑音条件である．N/A については，Fig. 8 と同様である．図より，提案手法（動的な自動閾値）は提案手法（最適な固定閾値）と同等の実発話率を得ていることが分かる．

これらの結果から，提案する自動閾値設定手法の有効性が確認された．

6　おわりに

本稿では，予測認識率が低い場合にユーザに発話音量を大きくするように通知することにより，認識不可となる頻度を低くし，かつ体感認識率を高く保つ手法を提案した．また，ユーザの要望や雑音環境に合わせて閾値を自動的に設定する手法を提案した．実験により，体感認識率を高く保ったまま，実発話率を改善できることが確認された．また、閾値を自動的に設定する手法の有効性が確認された．本稿の実験はシミュレーションであるため，今後は実際にシステムを実装し被験者実験を行う必要がある．

参考文献

[1] E. Morishita *et al.* , “Performance Estimation of Noisy Speech Recognition Based on Short-Term Noise Characteristics,” Proc. TJASSST ’11, pp. 1–4, Nov. 2011.

Table 1: Optimal fixed threshold for car noise and hall noise.

	走行自動車内			ホール		
	−10dB	0dB	10dB	−10dB	0dB	10dB
閾値 A	79	90	90	N/A	N/A	80
閾値 B	45	50	86	N/A	N/A	23

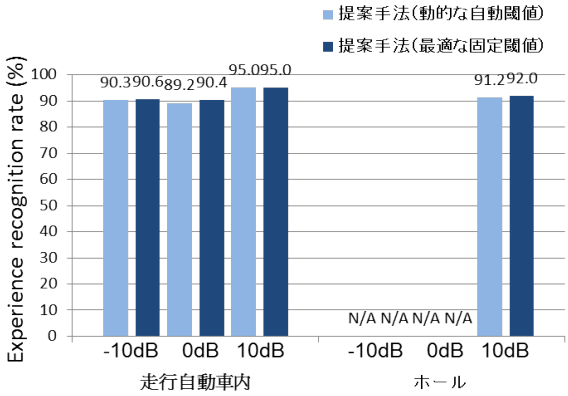

Fig. 8: Experience recognition rate using automatically set threshold and optimal fixed threshold.

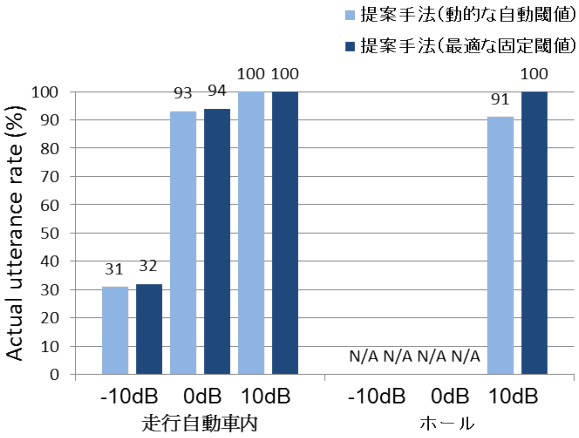

Fig. 9: Actual utterance rate using automatically set threshold and optimal fixed threshold.

[2] 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推定の検討,” 音講論, pp. 151–152, Mar. 2012.

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

[4] 東北大 - 松下 単語音声データベース, http://research.nii.ac.jp/src/TMW.html .

[5] 電子協騒音データベース, http://research.nii.ac.jp/src/JEIDA-NOISE.html .

[6] 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.

[7] 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.

[8] 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.