シミュレーションに基づく騒音下音声認識システム評価におけるロンバード効果の影響の検証 - 複数の認識タスク，騒音レベルに対する評価 - *

◎小川哲司, 倉持公壮, 小林哲則 (早稲田大学)

1 はじめに

シミュレーションに基づく騒音下音声認識システム評価におけるロンバード効果の影響を，複数の騒音レベル，認識タスク，認識システムに対して検証する．

騒音環境下において音声認識システムの評価を行う場合，部屋の特性，雑音の性質，話者の特性など，様々な要因の積で決まる状況ごとの音声を用いて認識実験を行うのが理想的である．しかし，それには膨大なテストデータの収集を必要とすることから，現実的な実験とは言えない．このため，通常は，音声認識に影響を与える要因を個別に捉え，その影響が独立に作用するものとして，それらの影響下での音声試験データを合成し実験に用いる．例えば，部屋の特性についてはインパルス応答をドライソースに畳み込むことで模擬し，騒音については独立に収集した騒音データを加算することで模擬する．しかし，騒音環境下において発話した場合，ロンバード効果と呼ばれる現象が生じ，声が大きくなり音高が上昇するなど，騒音の無い環境で発話した音声とは音声の特性が明らかに異なることが知られている [1, 2, 3, 4]．つまり，上述の方法では，騒音下で発話された音声を十分に模擬できている保証はない．

我々は，上述したシミュレーションに基づいて合成された音声を用いて音声認識性能を評価することの妥当性について実験的検証を行っている．特に，ロンバード効果による音響特性の変化が音声認識性能に及ぼす影響に焦点を当てて調査を行い，発話者に対し騒音を提示したときのドライソースを用いれば，騒音下での発話の認識性能を良好に模擬できる，騒音提示の音声認識性能への影響としては，ロンバード効果によりスペクトルが変形することよりも，発話のパワーが増大することで SNR が高くなることの方が重要な要因である，などの知見を得た [5]．しかし，これは特定の騒音レベル，認識タスクにおいて得られた知見である．

そこで本稿では，異なる騒音レベル，連続音声認識，孤立単語音声認識という異なる認識タスク，音響モデルの騒音適応の有無という異なる認識システムに対して，シミュレーションに基づく騒音下音声認識システム評価の妥当性の検証を行う．同時に，妥当なシミュレーションを行うための要件が，このような認識条件の違いにより受ける影響に関しても調査を行う．

以下，2 では収録した音声データとシミュレーションの方法について述べる．3 では音声認識実験を行い，音声認識性能の観点から，シミュレーションの可能性について検証を行う．最後に，4 でまとめとする．

2 シミュレーション

2.1 音声収録

ロンバード効果による音響特性の変化が音声認識性能に及ぼす影響を調査するために，一般的なオフィス環境において，下記の 3 通りの音声を収録した．

- 無騒音環境下 (30dB(A) 程度の背景雑音が存在) で発話されたロンバード効果の影響を含まない通常の発話音声 (CLEAN)

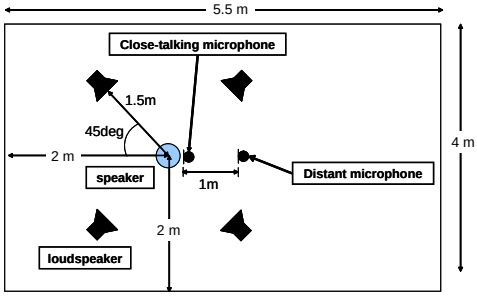

Fig. 1 The arrangement of a speaker and recording devices.

- 無騒音環境下で発話されたロンバード効果の影響を含む発話音声 (L-CLEAN)
- 騒音環境下で発話されたロンバード効果の影響を含む発話音声 (L-NOISY)

このとき，音声を収録した部屋におけるマイクロホン，スピーカ，発話者の配置を Fig.1 に示す．発話者は男性話者 10 人で，発話内容は，各話者につき，新聞記事読み上げ音声 50 発話と，音素連鎖バランス単語 100 発話である．このとき，各発話に対し，発話者の口元に設置された指向性の接話型マイクロホンと，発話者から 1 m 離れた位置に設置された無指向性の遠隔マイクロホンにより同時に収録を行った．CLEAN は，騒音が無い状況下で発話した音声を収録したものである．このとき，部屋には 30dB(A) 程度の背景雑音が存在する．L-NOISY は，4 台のスピーカから，電子協騒音データベース [6] に含まれる駅のコンコースの騒音を，発話者の耳元で 60dB(A) もしくは 70dB(A) になるように流し，この際の発話音声を収録することで得た．L-CLEAN の音声を収録する方法は次の通りである．まず，発話者の代わりにダミーヘッドを設置し，L-NOISY で用いたものと同一種類・大きさの騒音に対し，ダミーヘッドに設置されたマイクロホンによりバイノーラル録音を行った．このようにして収録された騒音をオープンエアのヘッドホンを通じて聞きながら発声した音声を，擬似的に騒音環境下で発話した音声として近似した．このとき，ヘッドホンを装着することで，発話者自身の声が抑圧されて聞こえる可能性がある．そこで，騒音環境下で発話することの自然性を考慮するために，自身の発声した音声を適切な音量に調整してヘッドホンよりフィードバックする必要がある．また，ヘッドホンから洩れた騒音がマイクロホンに入力される可能性についても検証する必要がある．前者に関しては，ダミーヘッドにヘッドホンを装着した場合としない場合の音声を収録したところ，その音圧差は無視できる程度に小さく，また，後者に関しても，ヘッドホンから洩れた音声のうちマイクロホンに入力されたものは無視できるほど小さいことがわかった．そこで，L-CLEAN の収録に際しては，発話者の音声のヘッドホンへのフィードバック，および，マイク入力に対するヘッドホンから洩れた音声の補正は行わない．この

*Influences of Lombard effect on simulation-based assessments of noisy speech recognition for various recognition tasks and noise levels. by T. Ogawa, K. Kuramochi, and T. Kobayashi (Waseda University)

Table 1 Evaluation items. "*" expresses computing a convolution. x denotes a noise level (in dB(A)). 60 and 70 are both applied to x in the present paper.

	原音声	発声様式	発話音声 (遠隔マイク)	騒音 (遠隔マイク)	SNR
LNx-D	L-NOISY	Lombard	直接入力	直接入力	——
Cx-DSN	CLEAN	neutral	ドライソース * インパルス応答	重畳	——
LCx-DSN	L-CLEAN	Lombard	ドライソース * インパルス応答	重畳	——
Cx-DSN-C	CLEAN	neutral	ドライソース * インパルス応答	重畳	= LCx-DSN

ようにして収録された 3 種類の音声を基に，Table 1 に示した 4 種類の音声を作成した．

LNx-D は，xdB(A) の騒音環境下における発話を遠隔のマイクロホンで直接収録した音声であり，ロンバード効果の影響が加味されている．Cx-DSN は，騒音環境下での発話をシミュレートした音声であり，発声様式は通常の発声である．これは，無騒音環境下で通常発声された (neutral) 音声に環境のインパルス応答を畳み込み，さらに xdB(A) の騒音を重畳することで得た，LCx-DSN は，同様に騒音環境下での発話をシミュレートした音声であるが，発話様式はロンバード発声である．これは，無騒音環境下でロンバード発声された (Lombard) 音声に環境のインパルス応答を畳み込み，xdB(A) の騒音を重畳することで得た，また，Cx-DSN-C は，Cx-DSN と同様の方法でシミュレートされた音声であるが，LCx-DSN と SNR が同一になるように，Cx-DSN における音声信号のパワーを調整することで得た．

2.2 妥当性の評価方法

騒音環境下での発話音声をシミュレートする場合，無騒音環境下で収録したドライソースに，その環境のインパルス応答を畳み込み，騒音を重畳することで，遠隔発話を近似した．このようなシミュレーションでは，ドライソースはロンバード効果が加味されていない通常発声であることがほとんどである．そこで，実際に騒音環境下で発話した音声 LNx-D の認識性能と，通常発声のドライソースに基づくシミュレーションにより合成された騒音下発話音声 Cx-DSN の認識性能，ロンバード発声のドライソースに基づくシミュレーションにより合成された騒音下発話音声 LCx-DSN の認識性能を直接比較することで，シミュレーションに基づく音声認識性能の妥当性を検証した．このとき，シミュレーションにより合成された騒音下発話音声の認識性能が，実音声 LNx-D と同等の性能を与えれば，妥当なシミュレーションであるとみなした．

2.3 シミュレーション方法

2.3.1 伝達関数測定

発話者から遠隔マイクロホンまでの伝達特性を time stretched pulse (TSP) 法によりインパルス応答として求めた．Fig.1 における発話者の位置にスピーカを配置し，TSP を出力した．TSP は，サンプリング周波数 32kHz，TSP 長 131072 ポイントの TSP up タイプを用い，8 回の同期加算を行った．得られたインパルス応答は，部屋の残響時間が 240ms であることから，残響時間が 240ms になるように切り出して用いた．また，このとき得られたインパルス応答には，スピーカの特性が含まれる．

2.3.2 収録機器の周波数特性の補正

無騒音環境下での近接発話音声を収録するにあたっては，SNR を向上させるために指向性ダイナミックマイクロホンを用いた．一般的に，指向性マイクロホンは，遠隔マイクロホンとして用いた無指向性マイクロホンに比べて平坦な周波数特性が得られない．そこで，発話者位置から近接マイクロホンへのインパルス応答を TSP 法により計測し，その逆フィルタを設計した．さらに，得られた逆フィルタを同一のマイクロホンで収録された音声に畳み込むことで，周波数特性の乱れを補正した．この逆フィルタを畳み込むことにより，インパルス応答測定の際に TSP を出力するためのスピーカの特性も同時に補正することができる．ただし，この逆フィルタを畳み込むことで音声のパワーが変化しないように，逆フィルタの大きさを調整した．

2.3.3 環境の模擬とパワーの調整

2.3.2 で得られた，マイクロホンとスピーカの特性を補正したドライソースに対して，2.3.1 で得られたインパルス応答を畳み込むことで，遠隔マイクロホンに到達する発話音声を模擬した．このようにして模擬された音声のパワーを，遠隔マイクロホンから直接入力した音声のパワーと一致するように調整した．このとき，発話者の口の位置とインパルス応答収録の際に TSP を出力したスピーカの位置は，発話ごとに厳密には異なることを考慮し，インパルス応答に基づく環境のシミュレーションを妥当に行うために，このパワーの調整を発話ごとに行った [5].

2.3.4 騒音の重畳

2.3.3 で得られた，環境の影響が模擬された音声に対して，遠隔マイクロホンで実際に収録した騒音を重畳した．この騒音は，発話者の耳元で 60dB(A) もしくは 70dB(A) になるように調整した．また，無騒音環境下においても 30dB(A) 程度の背景雑音が存在するため，背景雑音を収録し，重畳を行った．

3 音声認識実験

3.1 実験概要

騒音下音声認識のシミュレーションに基づく評価の妥当性に関して検証を行うため，音声認識実験を行った．本稿では，発話者の耳元に提示する騒音レベル (60dB(A)，70dB(A))，認識タスク (連続音声認識，孤立単語音声認識)，認識システム (音響モデルの騒音適応あり，無し) が異なる場合に対して検証を行うため，これらの組み合わせとして得られる，Table 2 に示した 8 通りのシステムを用いて実験を行った．

3.2 実験条件

3.2.1 連続音声認識

音響特徴量には，MFCC 12 次元，ΔMFCC 12 次元，Δ パワーの計 25 次元のパラメータを用いた．分析条件は，Table 3 に示す通りである．

音響モデルは，ASJ-JNAS，ASJ-PB[7] より，ヘッドセットを用いて収録した男性 133 話者の発話音声 (計 20406 文) から学習した，状態数 5，ループ数 3 の left-to-right 型の状態共有トライフォン HMM である．このとき，状態数は 2000，混合正規分布の混合数は 16，分散行列は対角共分散とした．

Table 2　Experimental items.

	認識タスク	騒音適応	騒音レベル
1)	連続音声認識	—	60dB(A)
2)	連続音声認識	—	70dB(A)
3)	連続音声認識	MLLR	60dB(A)
4)	連続音声認識	MLLR	70dB(A)
5)	孤立単語音声認識	—	60dB(A)
6)	孤立単語音声認識	—	70dB(A)
7)	孤立単語音声認識	MLLR	60dB(A)
8)	孤立単語音声認識	MLLR	70dB(A)

Table 3　Setup for acoustic feature extraction.

サンプリング周波数	16 kHz
フレーム長	25 ms
フレーム周期	10 ms
分析窓	ハミング窓
プリエンファシス	$1 - 0.97z^{-1}$

この音響モデルの騒音に対する適応を行う．HMM の学習に用いた ASJ-PB のうち男性 95 話者により発話された計 475 発話に，評価時に用いる騒音と同一種類 (駅のコンコース) の騒音を重畳して得られた音声を用いて，MLLR による適応を行った．MLLR の回帰クラス数は 4 とした．このとき，騒音を重畳する際の SN 比を 4 通り (5dB，10dB，15dB，20dB) に変化させて実験を行った．

言語モデルは，語彙数 2 万語彙の辞書を用いて作成した．語彙は，毎日新聞の 1991 年から 1994 年の 4 年分の新聞記事から構成されている．

評価データには，Fig. 1 の部屋において，60dB(A)，70dB(A) の騒音が提示されている状況下で，ASJ-JNAS の評価セットのうち 50 文を男性話者 10 名が実際に発話した音声を用いた．また，認識エンジンは，当研究室で開発したワンパストライグラムデコーダ SKOOD[8] である．

3.2.2　孤立単語音声認識

音響特徴量には，3.2.1 で用いたものと同一の 25 次元のパラメータを用いた．

音響モデルには，3.2.1 で用いた読み上げ音声 20406 発話から学習した，モノフォン HMM を用いた．このとき，音素クラス数は 43 である．また，混合正規分布の混合数は 32 とした．

このモノフォン HMM に対し，MLLR による騒音適応を行った．適応データ作成の際に重畳する騒音の種類や SN 比，適応データ数，MLLR の回帰クラス数などの条件は，3.2.1 と同一である．

評価データは，ATR 音素連鎖バランス単語 216 単語のうち 100 単語を，男性話者 10 名が Fig. 1 の部屋において 60dB(A)，70dB(A) の騒音が提示されている状況下で実際に発話した音声を用いた．このとき，辞書に登録されている語彙数は 216 である．

3.3　実験結果

Table 1 の評価項目に対する，連続音声認識における単語正解精度を Fig. 2 に，孤立単語音声認識における単語正解率を Fig. 3 に示す．各評価項目において，左側のグラフと右側のグラフは，各々騒音適応を行っていない場合の認識システムによる認識性能，騒音適応を行った場合の認識システムによる認識性能を表している．このとき，適応を行ったシステムの性能に関しては，適応データ作成の際に騒音を重畳する SN 比 5dB，10dB，15dB，20dB ごとに適応および認識を行い，その平均値を示した．

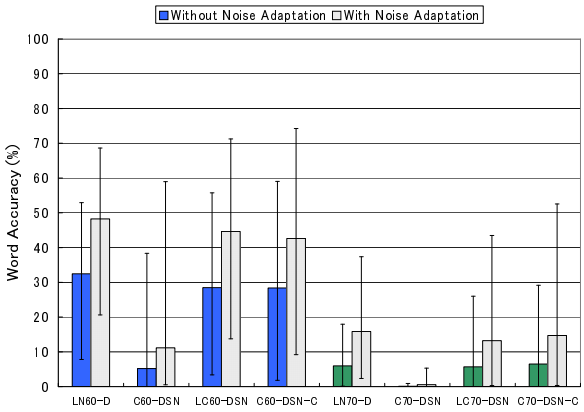

Fig. 2　Word accuracies for the real and the simulated noisy speeches under the continuous speech recognition condition. Each thick bar represents the average performance for 10 male speakers. The error bars represent the maximum and the minimum performance. The left side and the right side of each category represents the performance given by HMMs trained using clean speeches and that given by HMMs adapted to noisy speeches, respectively.

3.3.1　連続音声認識

通常発声のドライソースに基づき騒音下発話を模擬した Cx-DSN は，発話者が騒音を提示されることで生じるロンバード効果の影響を受けた実音声 LNx-D や合成された音声 LCx-DSN に比べて発話のパワーが小さく，良好な SNR を得られない．Fig. 2 からは，騒音レベルや認識システムにおける騒音適応の有無に依らず，Cx-DSN の性能は LNx-D の性能に対して著しく劣化していることが見て取れる．それに対し，LCx-DSN の性能は，騒音レベルや認識システムの騒音適応の有無に依らず，実音声 LNx-D とほぼ同一となっている．これより，ヘッドホンから実際と同一 (耳元で 60dB(A) もしくは 70dB(A)) の騒音を提示された状況で発話されたドライソースを用いてシミュレーションを行えば，騒音下での発話の認識性能を良好に模擬できることがわかる．

また，LCx-DSN の SNR に合わせるように Cx-DSN の発話のゲインを調整することで得られる Cx-DSN-C は，騒音レベルや認識システムにおける騒音適応の有無に依らず，LCx-DSN および LNx-D とほぼ同等の認識性能を与えた．このとき，Cx-DSN-C と Cx-DSN は共に通常発声であり，両者の認識性能の差は SNR による違いから生じている．また，Cx-DSN-C と LCx-DSN は同一の SNR であり，両者の認識性能の差は，通常発声とロンバード発声という異なる発声様式によるスペクトルの変形から生じている．したがって，連続音声認識においては，騒音を提示することの音声認識性能への影響としては，ロンバード効果によりスペクトルが変形することよりも，発話のパワーが増大することで SNR が大きくなることの方が重要な要因であり，発話のゲインを調整すれば，通常発声のドライソースを用いても，ロンバード発声のドライソースを用いても，妥当なシミュレーションを行うことが可能であると言える．これは，音声認識性能という観点からは，実際の騒音下発話音声の SNR を正確に推定することができれば，ロンバード発声のドライソースを用いること無しに妥当なシミュレーションを行えることを示唆している．

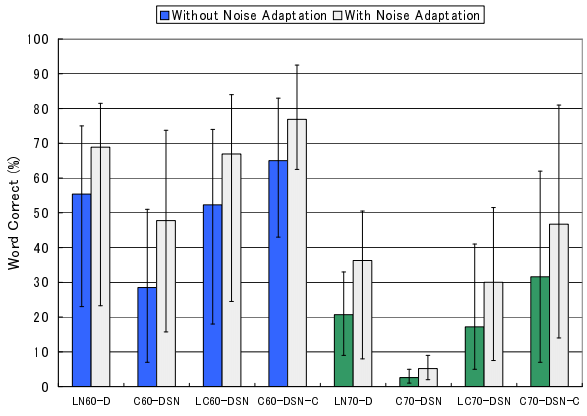

Fig. 3 Word corrects for the real and the simulated noisy speeches under the isolated spoken word recognition condition. Each thick bar represents the average performance for 10 male speakers. The error bars represent the maximum and the minimum performance. The left side and the right side of each category represents the performance given by HMMs trained using clean speeches and that given by HMMs adapted to noisy speeches, respectively.

3.4 孤立単語音声認識

Fig. 3 より，騒音レベル，認識システムにおける騒音適応の有無に依らず，通常発声のドライソースに基づき騒音下発話を模擬した Cx-DSN の性能は，騒音下で実際に発話した LNx-D の性能に対して著しく劣化しているのに対し，ロンバード発声のドライソースに基づき騒音下発話を模擬した LCx-DSN の性能は，LNx-D とほぼ同一の性能となった．したがって，連続音声認識実験の場合と同様，騒音を提示したときのドライソースを用いてシミュレーションを行えば，孤立単語音声認識においても騒音下での発話の認識性能を良好に模擬できることがわかる．

一方，Cx-DSN の SNR を LCx-DSN に合わせるようにゲイン調整を行った Cx-DSN-C の性能は，連続音声認識実験の場合とは傾向が異なり，LNx-D や，LCx-DSN の性能と同等とは言えない．実際，Cx-DSN-C は，LNx-D と比較して，騒音レベル，騒音適応の有無に関わらず，10 ポイント程度高い性能を与えている．これは，音響モデルが通常発声の音声を用いて学習されているため，ロンバード発声の音声よりも通常発声の音声の方が音響モデルとのミスマッチが小さいためと考えられる．連続音声認識においては，同一の SNR になるように発話音声のゲインを調節すれば，通常発声とロンバード発声という発声様式の違いによるスペクトル変形の影響を無視しても，妥当なシミュレーションが可能であった．これは，連続音声認識では，音響モデルの他に言語モデルも用いて探索を行うことで，音響的なスペクトル変形に起因する音響モデルの尤度の差が，言語モデルの尤度によって吸収されている可能性がある．しかし、孤立単語音声認識においては，音響モデルのみを用いて探索を行っていることから，通常発声とロンバード発声のスペクトルの違いによる音響尤度の差が認識性能に及ぼす影響を無視することができず，同一の SNR であっても，発声様式の違いにより認識性能は大きく異なる．したがって，孤立単語音声認識においては，通常発声の音声から学習された音響モデルを用い，通常発声のドライソースに基づくシミュレーションを行う場合，例え実際の騒音下発話の SNR を正確に推定することができたとしても，騒音下発話音声の認識性能を妥当に模擬することはできない．孤立単語音声認識において，通常発声されたドライソースに基づいて妥当なシミュレーションを行うためには，音響モデルのロンバード発声に対する適応などを施す必要があると考えられる．

4 まとめ

騒音環境下におけるシミュレーションに基づく評価実験の妥当性を，ロンバード効果による音響特性の変化が音声認識性能に与える影響に焦点を当てて検証した．その結果，60dB(A)，70dB(A) という騒音レベルの違い，連続音声認識，孤立単語認識という認識タスクの違い，騒音適応の有無という認識システムの違いに依らず，ヘッドホンから騒音を提示された状態で発話したドライソースを用いれば，騒音下での発話の認識性能を良好に模擬できることがわかった．

SNR を同一にしたときの，通常発声，ロンバード発声という発声様式の違いが認識性能に及ぼす影響に関しては，連続音声認識と孤立単語音声認識において異なる傾向が見られた．連続音声認識タスクにおいては，騒音レベル，認識システムの違いに依らず，発声様式の違いにより生じるスペクトル変形の認識性能への影響は無視できるほど小さく，発話のゲインを調整することができれば，通常発声のドライソースを用いても騒音下発話の認識性能を模擬できることがわかった．一方，孤立単語音声認識タスクにおいては，発声様式の違いによるスペクトル変形の認識性能への影響は，騒音レベル，認識システムの違いに依らず，無視できないほど大きく，同一の SNR になるように発話のゲインを調整しても，両者は同等の性能を与えないことがわかった．つまり，孤立単語音声認識実験において，通常発声の音声から学習した音響モデルを用いて妥当なシミュレーションを行うためには，ロンバード発声のドライソースに基づくシミュレーションを行う必要があることがわかった．

参考文献

[1] J.-C. Junqua, “The Influence of Acoustics on Speech Production: A Noise-Induced Stress Phenomenon Known as the Lombard Reflex,” Speech Comm., vol.20, pp.13-22, Nov. 1996.
[2] J. H. L. Hansen, “Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition,” Speech Comm., vol.20, pp.151-170, Nov. 1996.
[3] A. Wakao *et.al.*, “Variabitily of Lombard effects under different noise conditions,” Proc. ICSLP, vol.4, pp.2009-2012, 1996.
[4] H. Boril *et.al.*, “Design and Collection of Czech Lombard Speech Database,” Proc. Interspeech, pp.1577-1580, Sept. 2005.
[5] T. Ogawa *et.al.*, “Adequacy Analysis of Simulation-Based Assessment of Speech Recognition System,” Proc. ICASSP, vol.4, pp.1153-1156, April 2007.
[6] 電子協騒音データーベース, http://www.milab.is.tsukuba.ac.jp/corpus/noise_db.html
[7] K. Itou *et.al.*, “The Design of the Newspaper-Based Japanese Large Vocabulary Continuous Speech Recognition Corpus,” Proc. ICSLP, pp.3261-3264, Nov.1998.
[8] 柴田, 小林, “ワンパストライグラムデコーダにおける単語履歴の束ね処理に関する検討,” 音講論集, pp.151-152, Sept. 2002.