音声合成を用いた音声認識性能予測
– 残響と騒音が存在する環境での評価 – *

◎太刀岡勇気, 堀井昭男, 岩崎知弘 (三菱電機・情報総研),
斉藤辰彦, 能勢隆, 小林隆夫 (東工大)

1 はじめに

音声認識の実用化には、多様な話者・語彙での評価が必要であるが、発話の収集に多大なコストを要する。これに対し、合成音声により認識性能を予測する方法の有効性が示されている [1, 2]。ただし、これらの検討は残響・騒音のないクリーン環境での評価であり、実際の残響・騒音環境でも認識性能予測が有効であるかは明らかでない。そこで本報では残響・騒音環境で自然音声と合成音声の認識率の比較を行い、認識性能予測が有効であるか検討する。

2 音声合成を用いた音声認識性能予測

2.1 音声合成法

本報では、評価対象の目的話者の音声から HMM を学習し、音声を合成する HMM 音声合成法 [3] を用いた。これにより、任意の発話の評価が可能となる。各話者 (男女各 43 名)289 発話の自然音声から、メルケプストラム (0–24 次) および基本周波数とそれらの $\Delta, \Delta\Delta$ の 78 次元の特徴量を抽出し、特定話者モデルを学習した。音響モデルは 5 状態 left-to-right で状態継続長を明示的にモデル化した隠れセミマルコフモデルを用い、学習時のコンテキストは前後の音韻環境のみを考慮した。

2.2 実験条件

評価には、音声合成のモデル学習に使用した発声とは異なる 235 前後の住所、施設名の発声を使用した。なお発話者により内容は若干異なる。認識対象語彙は評価に用いた全 438 の住所、施設名とし、孤立単語認識を行った。認識率は単語単位で算出した。評価用の自然音声および合成音声は、残響下音声認識評価用データベース CENSREC-4[4] の 8 種の環境 (Fig. 2 凡例参照) のインパルス応答を畳み込み、騒音を SNR 15, 20, 25, 30 dB で加え作成した。

音響モデルは残響付加後に騒音を SNR 18, 24 dB で加えたデータから学習した (N-none_R-18-24[5])。音素片を基本単位とする混合数 8 の混合分布とし、MFCC (0–16 次) とその $\Delta, \Delta\Delta$ の 51 次元の特徴量を用いた。

Table 1 Correlation coefficients and RMSEs of average recognition rates between natural and synthetic speech samples in different environments.

SNR	30 dB	25 dB	20 dB	15 dB
Correlation	0.99	0.99	0.98	0.97
RMSE [%]	0.2	0.4	1.0	7.4

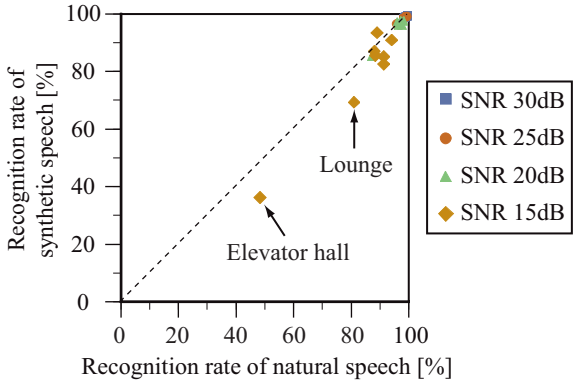

Fig. 1 Average recognition rates in respective environments and SNRs.

3 結果と考察

3.1 環境に対する認識性能予測

SNR 別の自然音声と合成音声の平均認識率の環境変化に関する相関係数、RMSE を Table 1 に示す。Fig. 1 は認識率の相関図である。このように SNR によらず、相関は高い。SNR 15 dB と SNR が低い場合は RMSE が大きいが、自然音声の認識率の低い Elevator hall で合成音声の認識率も低く、認識しづらい環境の発見に有効である。

3.2 話者に対する認識性能予測

各環境、SNR で話者ごとの認識率を自然音声と合成音声で比較した。相関と RMSE を Fig. 2 (i) に示す。SNR 30 dB では認識率がほぼ 100 %のため相関が低い環境が見られるが、SNR 25 dB 以下では 0.7 程度以上の相関はあり、認識率予測に有効である。SNR 15 dB での相関の高い Japanese room と相関の低い Lounge の相関図を Fig. 3 に示す。Lounge では合成音声の認識率が高/低に 2 極化している。

* Performance prediction of speech recognition using speech synthesis—Evaluation under reverberant and noisy environments, by TACHIOKA, Yuuki, HORII, Akio, IWASAKI, Tomohiro (Mitsubishi Electric Corp.), SAITO, Tatsuhiko, NOSE, Takashi, and KOBAYASHI, Takao (Tokyo Institute of Technology).

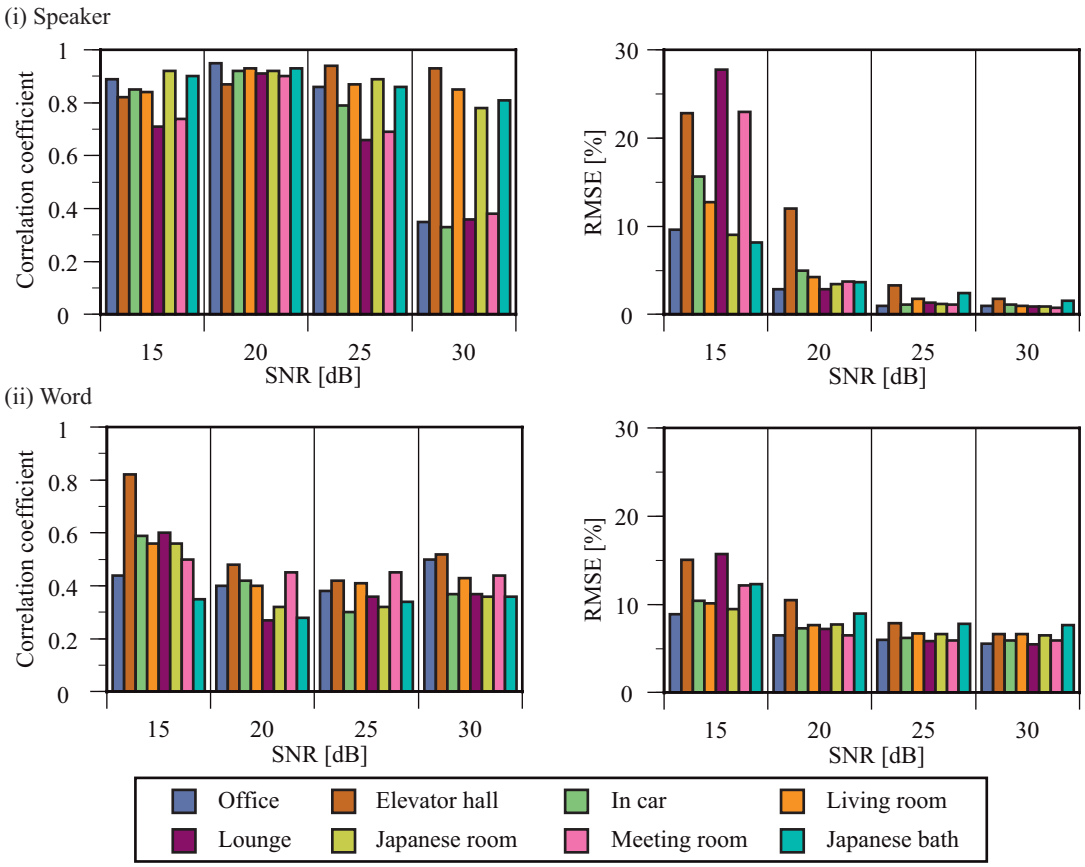

Fig. 2　Correlation coefficients and RMSEs of average recognition rates between natural and synthetic speech samples with respect to speakers and words.

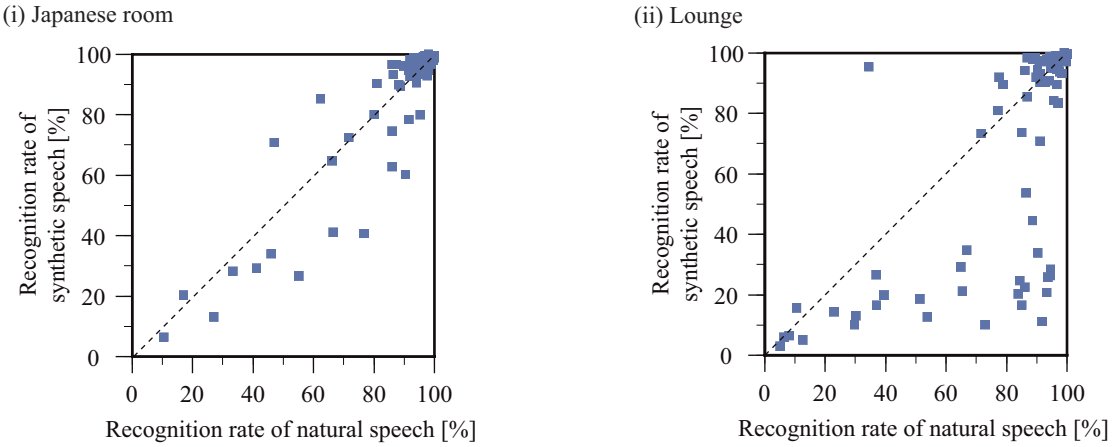

Fig. 3　Average recognition rate of respective speakers in different environments (SNR 15 dB).

3.3　単語に対する認識性能予測

単語ごとの認識率の相関と RMSE を Fig. 2 (ii) に示す。相関は 0.4 程度とあまり高くない。これは既報 [2] と同傾向で、話者ごとの平均的な発話様式の予測より、単語ごとの発話速度、音素の状態継続長等の予測が難しいためと考えられる。

4　まとめ

音声合成による認識率予測法を残響と騒音が存在する環境に適用した。本手法は、認識しづらい環境の発見に有効であった。また話者ごとの自然音声と合成音声の認識率の相関は高く、認識率の低い話者の発見に有効であった。今後の課題は、単語ごとの推定精度の改善である。

参考文献

[1] 寺嶌他, 電学論 C(130-4), pp.557-564, 2010.
[2] 斉藤他, 音講論 (春), pp.149-150, 2011.
[3] 吉村他, 信学論 J83-D-II(11), pp.2099-2107, 2000.
[4] http://research.nii.ac.jp/src/list/index.html
[5] 太刀岡他, 音講論 (秋), pp.17-20, 2010.