雑音環境下での非可聴つぶやき強調システムにおける目標音声の評価*

☆鶴田さくら，田中宏, 戸田智基, Graham Neubig, Sakriani Sakti, 中村哲（奈良先端大）

1 はじめに

非可聴つぶやき（Non-Audible Murmur: NAM）は，NAM マイクと呼ばれる専用の体表密着型マイクを用いて，体表から直接収録される肉伝導音声のひとつである [1]．NAM は秘匿性の高い発話を可能とするため，サイレント音声通話への応用が期待されている [2]．しかし，その音響特徴量は，通常の空気伝導音声のものとは大きく異なるため，明瞭性及び自然性が大きく劣化する．この問題に対処するため，統計的手法 [3][4] に基づき NAM から通常音声及びささやき声へと変換する NAM 強調法が提案されている [5]．これまでに，遮音室のような静環境下においては，NAM の品質を改善可能であること，また，通常音声よりもささやき声への変換の方が有効であることが報告されている．しかし，実環境下では，受聴時及び発声時に外部雑音の影響を受けるため，同様の結果が得られるとは限らない．

本研究では，NAM 強調技術の実環境への適用を目指して，受聴者が雑音環境下にいる状況を想定した際に，変換に用いる最適な目標音声に対する評価を行う．なお，発話者に関しては，静穏下にいる状況を想定する．

2 統計的手法に基づく NAM 強調法

統計的手法に基づく NAM から通常音声への変換（NAM2SP）[5] では，NAM のスペクトル特徴量（メルケプストラムセグメント）を入力とし，通常音声のスペクトル特徴量（メルケプストラム），有声無声情報を含む F_0，及び非周期成分の各々を出力とする計 3 つの GMM を学習し，最尤系列変換法 [4] に基づき各々変換を行う．変換後の F_0 および非周期成分を用いて混合励振源モデル [6] により音源信号を生成する．その後，音源信号に対して，変換後のスペクトル特徴量を畳み込むことで，強調音声を得る．本手法により得られた強調音声は，NAM に比べて明瞭性が改善されており，通常音声と類似した音質を持つ．一方で，無声音である NAM のスペクトル特徴量から通常音声の F_0 パターンを予測することは本質的に困難である．結果，生成される強調音声は不自然な抑揚を持つ．

この問題を回避するため，NAM のスペクトル特徴量から，無声音であるささやき声への変換（NAM2WH）[5] が提案されている．ささやき声は，NAM と比較して，馴染みのある空気伝導音声であるため，高い明瞭性及び自然性を有する．統計的手法に基づく NAM からささやき声への変換では，NAM のスペクトル特徴量（メルケプストラムセグメント）からささやき声のスペクトル特徴量（メルケプストラム）を変換する．雑音源に対して，変換後のスペクトル特徴量を畳み込むことで，強調音声を得る．本手法では，通常音声への変換とは異なり，F_0 パターン推定処理を回避することができる．また，無声音同士の変換であるため，通常音声への変換よりも，高いスペクトル特徴量変換精度が得られる．結果，NAM と比べて，強調音声の自然性および明瞭性は大幅に改善され，通常音声への変換よりも高い性能が得られる．

Table 1 変換音声に使用する特徴量

	スペクトル	音源
CVSP	NAM2SP	NAM2SP
CVWH	NAM2WH	Noise
CVWH+CVSRC	NAM2WH	NAM2SP
WH+CVSRC	WH	NAM2SP

3 最適な変換目標音声の調査

サイレント音声通話における NAM 強調処理では，明瞭性や聞き取りやすさの改善が重要となる．これまでに，遮音室のような静環境下においては，自然音声としては，通常音声の方がささやき声よりも明瞭性が高いこと，一方で，変換音声としては，変換ささやき声の方が変換通常音声よりも明瞭性が高いことが報告されている．これに対して，雑音環境下においては，雑音と区別がしやすい音声へと変換することが有効であると予想される．

本研究では，実環境下での受聴を想定し，雑音環境下において最適な目標音声の調査を行う．調査対象とする音声は，従来用いられていた通常音声（SP），変換通常音声（CVSP），ささやき声（WH），変換ささやき声（CVWH）に加え，通常音声とささやき声の中間の音声として，ささやき声スペクトルに有声音源を付与した以下の 2 つも用いる．

- 変換ささやき声スペクトル＋変換音源（CVWH+CVSRC）
- ささやき声スペクトル＋変換音源（WH+CVSRC）

変換ささやき声スペクトル＋変換音源は，NAM と通常音声との間で学習された GMM を用いて得られる変換音源に対して，NAM とささやき声との間で学習された GMM を用いて得られる変換スペクトル特徴量を畳み込むことにより得られる強調音声である．また，この強調音声の上限値として，理想的なスペクトル特徴量変換が行われた際を想定し，ささやき声スペクトル＋変換音源を用いる．これは，上記の変換音源に対し，自然なささやき声から抽出されたスペクトル特徴量を畳み込むことで得られる．なお，NAM マイクのみでなく空気伝導マイクも用いた収録を行うことで，実際にこの強調音声を得るシステムを構築することができる．表 1 に，各変換音声に使用する特徴量について示す．

4 実験的評価

4.1 実験条件

学習データとして ATR 音素バランス文セット中の 50 文中 40 文を用い，評価データとして残りの 10 文を用いる．女性話者 1 名による NAM 収録を行う．

なお，本研究では，NAM とささやき声の中間程度の音量での発話を取り扱い，NAM マイクと空気伝導

*Evaluation of target speech for a nonaudible murmur enhancement system in noisy environments, by TSURUTA, Sakura, TANAKA, Kou, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, and NAKAMURA, Satoshi (Nara Institute of Science and Technology)

Table 2　統計的手法に基づく NAM の変換精度

	通常音声	ささやき声
Mel-cepstral distortion	5.9 dB	4.9 dB
Aperiodic distortion	5.2 dB	
U/V error rate	16.1 %	
F_0 correlation coefficient	0.54	

マイクを用いて体内伝導音声（以下，NAM）と空気伝導音声（以下，ささやき声）を同時に収録する．サンプリング周波数は 16 kHz とする．入力特徴量として，FFT 分析による 0〜24 次のメルケプストラムセグメント特徴量（前後 4 フレーム相当）を用いる．通常音声及びささやき声のスペクトル分析には STRAIGHT 分析 [7] を用いる．GMM の混合数は 32（スペクトル変換用），16（F_0 変換用），8（非周期成分変換用）とする．その際の推定精度を表 2 に示す．なお，メルケプストラム歪みは，0 次項を含まずに計算している．

主観評価実験として，3 節で説明した 6 手法の音声の聞き取りやすさに関して 5 段階オピニオン評定により評価を行う．音声の受聴は，携帯電話の使用状況を想定し，オープンイヤーのヘッドホンによる片側受聴とする．実験を行う環境は，雑音なしの静環境下を想定した遮音室（25 dB(A)），雑音環境下を想定して遮音室にてスピーカからオフィス雑音（50 dB(A)）及び人混み雑音（70 dB(A)）の各々を提示した場合の計 3 種類とする．なお，ヘッドフォンからの提示音声のレベルは 55 dB(A) である．

被験者は男性 8 名，女性 2 名で，1 人あたり各環境，音声につき 6 サンプル，計 108 サンプルを受聴する．

4.2　実験結果

図 1 に遮音室，オフィス雑音，人混み雑音のそれぞれの環境における聞き取りやすさに関する評価結果を示す．遮音室においては，通常音声（SP），変換通常音声（CVSP），ささやき声（WH），及び変換ささやき声（CVWH）の関係は，文献 [6] で示されている明瞭性の結果と類似しており，通常音声，ささやき声，変換ささやき声，変換通常音声の順番で聞き取りやすいことが分かる．これらの音声に関しては，雑音環境下では聞き取りやすさの低下が見られるが，オフィス雑音下においてはささやき声（WH）と変換ささやき声（CVWH）の差は無くなり，人混み雑音下においては逆に変換ささやき声（CVWH）の方がささやき声（WH）を上回る．この原因については詳細に検討する必要がある．また，ささやき声／変換ささやき声に対する変換音源の付与処理（+CVSRC）に関しては，遮音室およびオフィス雑音下においては，劣化を引き起こす傾向が見られる．一方で，人混み雑音下においては，ささやき声（WH）に対して変換音源を付与することで，大幅な改善が見られる．このことから，雑音が大きな環境下においては，有声音源は聞き取りやすさを向上させる要因になると推測できる．しかしながら，変換ささやき声（CVWH）に対しては，その改善効果は小さくなる傾向が見られる．また，人混み雑音下での変換ささやき声（CVWH）と変換ささやき声スペクトル＋変換音源（CVWH+CVSRC），変換ささやき声スペクトル＋変換音源とささやき声スペクトル＋変換音源（WH+CVSRC），変換ささやき声とささやき声スペクトル＋変換音源で t 検定を行ったところ，統計的有意差は見られなかった．

以上より，静環境下および雑音環境下においても，

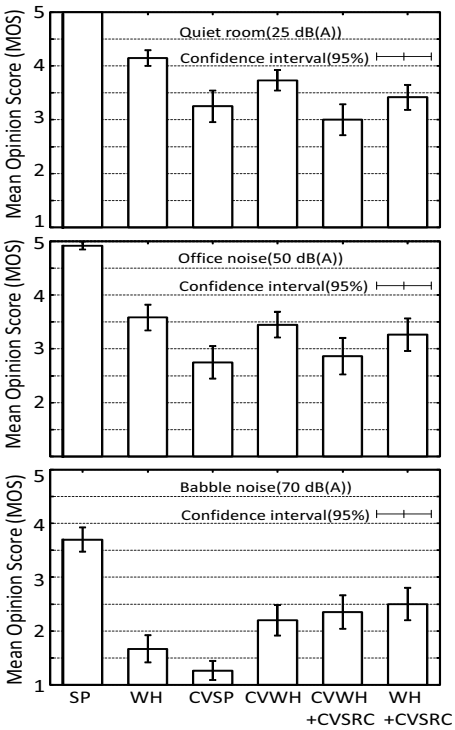

Fig. 1　主観評価実験結果（上図が遮音室 (25 dB(A))，中図がオフィス雑音 (50 dB(A))，下図が人混み雑音 (70 dB(A)) における評価結果）

変換ささやき声が有効であり，また変換音源を付与することでさらなる改善が得られる可能性があることが分かる．

5　おわりに

本稿では，NAM 強調技術の実環境への適用を目指し，雑音環境下における受聴を考慮した最適な変換目標音声の評価を行った．主観評価実験の結果から，静環境下および雑音環境下においても変換ささやき声が有効であること，また，外部雑音が大きな環境下においては変換音源の付与により聞き取りやすさを改善できる可能性があることが分かった．今後は，ささやき声へのスペクトル変換精度の改善に取り組むとともに，話し手側における外部雑音の影響について調査する．

謝辞　本研究の一部は，JSPS 科研費 22680016 の助成を受け実施したものである．

参考文献

[1] 中島 他，信学論，vol. 87，no. 9，pp. 1757-1764, 2004.
[2] B. Denby *et al., Speech Commun.*, vol. 52, no. 4, pp. 270-287, 2010.
[3] Y. Stylianou *et al., IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131-142, 1998.
[4] T. Toda *et al., IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222-2235, 2007.
[5] T. Toda *et al., IEEE Trans.* ASLP, Vol.20, No.9, pp.2505-2517, 2012.
[6] 大谷 他，信学論，vol.91，no.4，pp. 1082-1091, 2008.
[7] H. Kawahara *et al., Speech Commun.*, vol.27，no.3-4，pp. 187-207, 1999.