雑音環境下での受聴を想定した 非可聴つぶやき強調処理における変換音声有声化の効果*

☆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体内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，遮音室のような静穏環境下での受聴評価において，NAM 強調法により 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，受聴時の環境は静穏環境下に限定されないため，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の第一歩として，まずは，受聴者が雑音環境下，発話者が静穏環境下にいる状況を想定して，最適な変換目標音声の調査を行う．

2 統計的手法に基づく NAM 強調法

統計的手法に基づく NAM 強調法では，通常音声やささやき声といった自然音声を目標音声として，NAM の音響特徴量を目標音声の特徴量へと変換することで，NAM の品質を改善する．本手法は，学習処理と変換処理で構成される．学習処理では，NAM と目標音声の同一内容発話音声データに対してフレーム間の対応付けを行った結合ベクトルを用いて，NAM の音響特徴量と目標音声の音響特徴量の結合確率密度を混合正規分布モデル（GMM: Gaussian mixture model）でモデル化する．変換処理では，最尤系列変換法 [4] により，NAM の音響特徴量系列を目標音声の音響特徴量系列へと変換し，音声波形を合成することで，強調音声を得る．

3 雑音環境下での受聴に適した変換目標音声

NAM を用いてサイレント音声通話を行う上で，自然性の改善も必要ではあるが，明瞭性や聞き取りやすさの改善が最も重要となる．NAM は特別な発話様式であり，発話が周囲の迷惑となる図書館などの静穏環境下など，話者が NAM を使用し得る環境は比較的限られる．一方で，通話相手である受聴者の環境は限定されないため，雑音環境下で強調音声を受聴する状況が起こり得る．そのため，雑音環境下においても明瞭性の高い強調音声の実現が望まれる．

本研究では，雑音環境下での受聴を想定して，最適な目標音声の調査を行う．調査対象とする目標音声は，従来用いられていた NAM，通常音声，ささやき声に加え，有声化したささやき声及び電気音声（EL）の 2 つも新たに検討する．

3.1 有声化ささやき声への変換

雑音環境下において音声を受聴する際に，F_0 は一つの大きな手掛かりとなる．しかしながら，ささやき声は無声音であるため，雑音環境下での明瞭性が大幅に劣化する可能性がある．そこで，ささやき声を有声化した音声への変換を提案する．

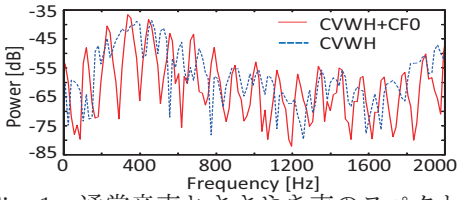

Fig. 1 通常音声とささやき声のスペクトル

有声化ささやき声へと変換するために，NAM のスペクトルセグメント特徴量から，ささやき声のスペクトル特徴量と通常音声の音源特徴量への変換を行う．音源特徴量として，連続 F_0 パターン（CF_0）[6] と非周期成分を用いる．これらの通常音声の音源特徴量を用いて混合励振源により音源波形を生成し，ささやき声のスペクトル特徴量を畳み込むことで，変換有声化ささやき声（CVWH+CF_0）を生成する．

図 1 に，変換ささやき声（CVWH）と CVWH+CF_0 のスペクトルの一例を示す．ここで，どちらの変換音声も波形パワーは等しいことに注意する．有声化することで，スペクトルは調波構造を持つため，調波成分に対応する周波数成分のパワーが大きくなることが分かる．このため，雑音環境下においては，有声化により，知覚に重要な周波数成分のパワーが相対的に大きくなるという効果が期待される．

3.2 電気音声への変換

CVWH+CF_0 は，無声音であるささやき声のスペクトルと有声音である通常音声の音源特徴量の組み合わせであるため，スペクトル特徴量と音源特徴量との間で有声無声情報の不一致が生じ，通常の音声生成過程では発声不可能な音声となる．聞きなれない音声となるため，十分な明瞭性が得られない可能性がある．

そこで，有声音であり，CVWH や CVWH+CF_0 と同様に有声無声情報を一切必要としない変換目標音声として，EL の使用を提案する．EL は，喉頭摘出者の代替発声器具の一つである電気式人工喉頭を用いて生成される音声であり，機械的に生成される音源信号を調音することで得られる有声音声である．例えば文献 [7] で報告されているとおり，EL は明瞭性が比較的高いことが既に知られている．ただし，EL は F_0 パターンが人工的であるため，自然性が大きく劣化する．そのため，提案法では，NAM のスペクトルセグメント特徴量から，EL のスペクトル特徴量と，通常音声の音源特徴量である CF_0 および非周期成分への変換を行うことで，より自然な F_0 パターンを持つ強調音声（CVEL+CF_0）を生成する．なお，EL として，NAM の発話者と同一話者が電気式人工喉頭を用いて発声したものを用いる．

4 実験的評価

4.1 実験条件

男性話者 1 名の通常音声（SP），ささやき声（WH），EL を空気伝導マイクを用いて収録する．また，同一話者の NAM を，NAM マイクと空気伝導マイクの両方を用いて同時に収録する．収録文は ATR 音素バランス文セット中の 50 文とする．また，親密度別単語了解度試験用音声データセット 2007[8] 中の親密度 1 の単語 8 セット（計 160 単語）も収録する．サンプリング周波数は 16 kHz とする．

NAM の音響特徴量として，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

---

*Voicing Effects in Statistical NAM Enhancement on Intelligibility of Converted Speech in Noisy Environments, by TSURUTA, Sakura, TANAKA, Kou, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, and NAKAMURA, Satoshi (Nara Institute of Science and Technology)

Table 1　統計的手法に基づく NAM の変換精度

	CVSP	CVWH	CVWH+CF_0	CVEL+CF_0
Mel-cepstral dist.	4.2 dB	4.5 dB	4.5 dB	5.4 dB
Aperiodic dist.	3.6 dB	N/A	3.6 dB	3.6 dB
U/V error	8.9%	N/A	13.5%	13.5%
F_0 correlation coef.	0.43	N/A	0.53	0.53

レーム相当）を用いる．通常音声及び EL のスペクトル分析には STRAIGHT 分析 [9] を用いる．学習データとして 40 文を用い，評価データとして残りの 10 文を用いる．GMM の混合数は 64（スペクトル変換用），32（F_0，CF_0 変換用），16（非周期成分変換用）とする．なお，文献 [5] とは異なり，NAM と通常音声においてフレーム間の対応付けを行う際には，空気伝導マイクで同期収録された NAM を用いる．強調音声として，従来の変換通常音声（CVSP）と変換ささやき声（CVWH），新たに提案する CVWH+CF_0 と CVEL+CF_0 を合成する．参考までに，各強調音声生成時における音響特徴量の推定精度を表 1 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，入力音声である NAM，目標音声である SP と WH と EL，強調音声である CVSP と CVWH と CVWH+CF_0 と CVEL+CF_0 の計 8 種類を評価する．まず，音声の聞き取りやすさに関する 5 段階オピニオン評定による評価を行う．受聴環境として，雑音なしの静穏環境とオフィス雑音を SNR=0 dB 及び-10 dB で重畳した場合の雑音環境の計 3 種類を想定する．被験者は 12 名で，1 人あたり各環境，音声につき 10 サンプル，計 240 サンプルを受聴する．

また，親密度 1 の単語セットを用いて明瞭性に関する書き取り試験も行う．評価音声として，NAM，CVSP，CVWH，CVWH+CF_0 を用いる．受聴環境は，雑音なしの静穏環境と，オフィス雑音を SN 比 0 dB で加えた雑音環境の計 2 種類とする．被験者は 4 名で，1 人あたり各環境，音声につき 20 サンプル，計 160 サンプルを受聴する．

4.2　実験結果

図 2 に各環境下における聞き取りやすさに関する評価結果を示す．自然音声については，全ての環境において SP が最も高く，NAM は最も低い．また，静穏環境下においては EL よりも WH の方が高いのに対し，雑音環境ではその差がほぼなくなる．この結果から，聞き取りやすさに関して，無声音は有声音よりも外部雑音の影響を受けやすいことが分かる．一方で，NAM と各種強調音声を比べると，全ての環境において NAM 強調法により NAM の聞き取りやすさを改善出来ていることが分かる．CVWH と CVWH+CF_0 を比べると，静穏環境下では CVWH の方が聞き取りやすいのに対し，外部雑音が大きくなるにつれて結果が逆転する傾向にある．このことから，ささやき声を有声化することで，外部雑音に対する頑健性を高めることが出来ると言える．一方で，CVEL+CF_0 は他の強調音声に比べて値が低い．これは，表 1 に示した通り，変換精度が低いのが原因だと考えられる．なお，CVSP は CVWH よりも同等か上回る値を示している．

表 2 に明瞭性に関する書き取り試験結果を示す．NAM 強調法により，NAM の明瞭性を改善出来ることが分かる．CVWH は雑音環境下において明瞭性がかなり劣化しているのに対し，CVSP と CVWH+CF_0 は外部雑音による明瞭性の劣化は小さい．このことから，雑音環境下において，有声音声への変換は無声音声への変換よりも有効であることが分かる．また，CVWH と CVWH+CF_0 の比較から，静穏環境下においても，ささやき声を有声化することで，明瞭性を改善出来る傾向が見られる．なお，文献 [5] の報告と異なり，CVSP が CVWH よりも明瞭性が高いという結果が得られている．この原因として，話者の違いや評価文の違い，さらには学習時におけるフレーム間対応付け処理の違いが考えられるため，さらなる調査が必要である．

以上より，NAM 強調法において，有声音声への変換処理を用いることで，受聴時における外部雑音の影響を低減させ，より明瞭で聞き取りやすい強調音声を得ることが出来ることが分かる．

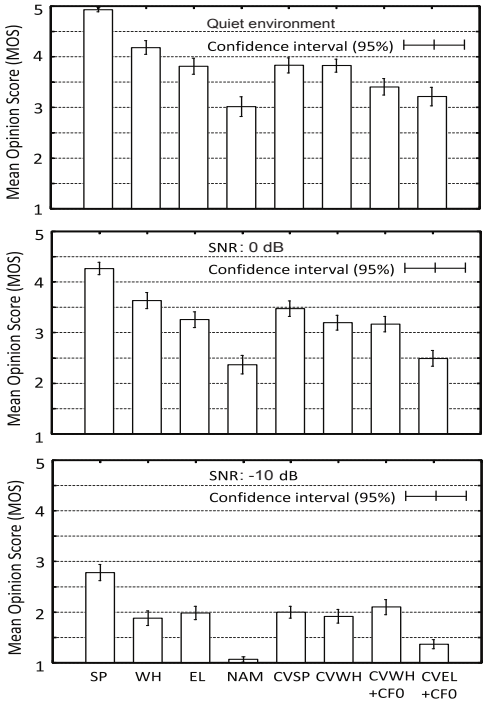

Fig. 2　主観評価実験結果（上図が静穏環境，中図が SNR = 0 dB，下図が SNR = -10 dB における評価結果）

Table 2　書き取り試験結果

	Mora correct rate (Quiet environment)	Mora correct rate (SNR = 0 dB)
NAM	55.0%	42.5%
CVSP	67.8%	63.1%
CVWH	61.6%	55.0%
CVWH+CF_0	65.3%	62.5%

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想定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，通常音声や有声化されたささやき声のような有声音は，ささやき声のような無声音に比べて，聞き取りやすさ及び明瞭性の面において，外部雑音に対して頑健であることが分かった．今後は，より明瞭性の高い変換目標音声のさらなる調査を行うとともに，雑音環境下における通常音声に対する明瞭性改善技術を NAM 強調システムに導入する．

謝辞　本研究の一部は，JSPS 科研費 26280060 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.*.SAP, vol. 6, no. 2, pp. 131-142, 1998.
[4] T. Toda *et al., IEEE Trans.*ASLP, 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] K. Tanaka *et al., IEICE Trans*, vol. E97-D, no. 6, pp. 1429-1437, 2014.
[7] T. Toda *et al., IEEE Trans.* ASLP, vol. 22, no. 1, pp. 172-183, 2014.
[8] 近藤 他，信学技報，WIT2007-62，pp. 43-48，2008.
[9] H. Kawahara *et al., Speech Commun.*, vol. 27, no. 3-4，pp. 187-207, 1999.