非可聴つぶやき強調音声の雑音環境下における明瞭性改善に関する検討*

☆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

秘匿性の高い音声コミュニケーションを実現する手法の一つとして，非可聴つぶやき（Non-Audible Murmur: NAM）[1] を用いたサイレント音声通話に関する研究が行われている．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ため，通常の空気伝導音声と比べて，明瞭性及び自然性が大きく劣化する．これに対し，統計的音声変換法 [3][4] に基づき，NAM をより自然な音声へと変換する NAM 強調技術が提案されている [5]．我々は，NAM 強調技術の実環境への適用を目指し，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，明瞭性に着目した最適な変換目標音声の調査を行ってきた [6]．その結果，1 名の話者を対象とした受聴評価において，雑音環境下においては，有声音への変換が効果的であることを示した．一方で，静穏環境下においても同様の傾向が見られており，無声音であるささやき声への変換の方が有声音である通常音声への変換よりも有効とする文献 [5] と異なる結果が得られていた．そのため，周囲の環境に応じて最適な変換目標音声が異なるか否かについては，未だ明らかにされていない．

本研究では，NAM の発声は個人差が大きい点 [7] に着目し，複数人話者の NAM に対して変換目標音声の評価を行う．評価結果から，静穏環境下における変換音声の明瞭性は話者に大きく依存すること，結果，静穏環境下における最適な目標音声は話者により異なること，また，雑音環境下においては話者に限らず有声化による明瞭性改善効果が得られることを示す．

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音響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．ここで，変換目標音声として用いる音声の種類に応じて，取り扱う音響特徴量は異なる．無声音であるささやき声への変換においては，スペクトル特徴量のみを推定するのに対し，有声音である通常音声への変換においては，スペクトル特徴量のみでなく F_0 や非周期成分といった音源特徴量も推定する．

文献 [5] では，女性話者 1 名を対象とした書き取り評価により，静穏環境下においては，変換通常音声（CVSP）よりも変換ささやき声（CVWH）の方が明瞭性が高いと報告されている．一方で，文献 [6] では，男性話者 1 名を対象とした書き取り評価により，静穏環境下においても CVSP の方が CVWH よりも明瞭性が高いという結果が報告されている．

3 雑音環境下での受聴を想定した NAM 強調法

NAM を用いてサイレント音声通話を行う上で，自然性よりも明瞭性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，静穏環境下のみでなく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，様々な環境下においても高い明瞭性が得られる強調音声の実現が望まれる．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り得る．CVWH は静穏環境下では高い明瞭性を持つとされているが，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高い．

雑音環境下において明瞭性の高い強調音声を得る手法として，我々はささやき声を有声化した音声への変換処理を提案している [6]．NAM からささやき声のスペクトル特徴量推定は，通常音声のスペクトル特徴量推定よりも精度が高い点に着目し，NAM 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を推定することで，有声化ささやき声への変換を行う．この時，音源特徴量として，連続 F_0 パターン（CF0）[8] と非周期成分を用いる．推定された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，別途推定された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+CF0）を生成する．

4 実験的評価

客観評価として，スペクトル特徴量変換精度の比較を行う．また，主観評価として，明瞭性に関する書き取り試験を行う．

4.1 実験条件

男性話者 2 名，女性話者 1 名の通常音声（SP），及びささやき声（WH）を収録する．さらに，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．男性話者 2 名をそれぞれ話者 A，話者 B，また女性話者を話者 C とする．客観評価に用いる学習データは，ATR 音素バランス文 A セット中の 40 文であり，評価データとして残り 10 文を用い，5 交差検定を行う．主観評価については，学習データとして ATR 音素バランス文セット中の 50 文を用い，評価データとして親密度別単語了解度試験用音声データセット 2007[9] 中の親密度 1 の単語 8 セット（計 160 単語）を用いる．サンプリング周波数は 16 kHz とする．

---

*Investigation on Intelligibility Improvements by Nonaudible Murmur Enhancement under Noisy Environments, by TSURUTA, Sakura, TANAKA, Kou, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, and NAKAMURA, Satoshi (Nara Institute of Science and Technology)

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係数に対するセグメント特徴量（前後 4 フレーム相当）を用いる．通常音声のスペクトル分析には STRAIGHT 分析 [10] 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．

4.2 客観評価実験

3 人の話者のスペクトル特徴量変換精度（メルケプストラム歪み）を表 1 に示す．スペクトル変換のための GMM として，通常音声への変換用を GMM_SP，ささやき声への変換用を GMM_WH と表記する．なお，メルケプストラム歪みは，0 次項を含まずに計算する．

男性話者 A に関して，GMM_SP は，GMM_WH よりも高い精度が得られており，文献 [5] に示されている 3 名の話者とは異なる傾向を示している．一方で，男性話者 B と女性話者 C に対しては，どちらの話者においても，GMM_WH の方が，GMM_SP よりも精度が高く，文献 [5] と同様の傾向が見られる．以上の結果から，通常音声のスペクトル特徴量への変換は，必ずしもささやき声のスペクトル特徴量への変換よりも精度が劣るわけでは無く，話者によってはささやき声の場合と同等以上の精度が得られることが分かる．

4.3 主観評価実験

明瞭性に関する単語書き取り試験を行う．使用する単語は，全て 4 モーラで構成されている．評価音声として，入力音声である NAM と，強調音声である CVSP と CVWH，及び CVWH+CF0 の計 4 種類を用いる．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 8 名で，1 人あたり各環境，話者，音声につき 20 サンプル，計 480 サンプルを受聴する．

表 2 に明瞭性に関する書き取り試験結果として，モーラ正解率を示す．静穏環境下において，話者 A に対しては，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．一方で，話者 B と話者 C については，文献 [5] と同様に，CVWH の方が CVSP よりも明瞭性が高い．平均的には，CVWH の方が CVSP よりも明瞭性が高い傾向が見られる．これらの傾向は，表 1 に示したスペクトル特徴量変換精度の傾向とも一致している．したがって，静穏環境下においては，必ずしも CVWH の方が CVSP よりも明瞭性が高くなるわけではなく，話者に依存することが分かる．また，CVWH と CVWH+CF_0 の比較から，静穏環境下においては，有声化により明瞭性が劣化することが分かる．なお，どの話者においても，NAM 強調法により NAM の明瞭性を改善出来ることが分かる．また，話者によって得られる改善効果が大きく異なることが分かる．

一方，雑音環境下においては，CVWH の明瞭性は大きく劣化する．特に，話者 B に関しては，明瞭性改善効果が得られていないことが分かる．一方で，有声音である CVSP と CVWH+CF0 は外部雑音による明瞭性劣化の影響が小さい．特に，CVWH+CF0 に関しては，全ての話者において，最も高い明瞭性が得られており，NAM の明瞭性を改善出来ることが分かる．

以上より，1) 受聴者が静音環境下にいる状況においては，CVWH による明瞭性改善効果が大きいが，話者によっては CVSP も同等以上の明瞭性改善効果が得られること，2) 受聴者が雑音環境下にいる状況においては，有声音への変換処理に基づく NAM 強調法を用いることで，受聴時における外部雑音の影響を低減させ，より明瞭性の高い強調音声を得ることが出来ることが分かる．

Table 1 話者毎のスペクトル特徴量変換精度

	GMM_SP	GMM_WH
Speaker A	4.2 dB	4.5 dB
Speaker B	5.2 dB	5.0 dB
Speaker C	5.7 dB	5.2 dB

Table 2 各話者毎のモーラ正解率（%）

	Mora correct rate(%) (Quiet environment) (SpeakerA/B/C : Avg.)	Mora correct rate(%) (SNR = 0 dB) (SpeakerA/B/C : Avg.)
NAM	35.0/ 20.3/ 26.9 : 27.4	23.1/ 21.3/ 23.8 : 22.7
CVSP	**70.0**/ 21.9/ 33.8 : 41.9	61.8/ 21.0/ 32.3 :38.3
CVWH	61.6/ **31.9**/ **40.0** : 45.1	59.6/ 20.3/ 30.3 : 36.8
CVWH+CF_0	60.6/ 25.3/38.4 : 41.6	**62.5**/ **26.1**/ **34.2** : 41.0

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，複数話者の NAM データを対象として，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，明瞭性において，通常音声や有声化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外部雑音に対して頑健であり，特に，有声化ささやき声への変換が有効であることを示した．また，静穏環境下においては，ささやき声への変換が有効であるが，話者によっては通常音声への変換も同等以上の明瞭性改善効果が得られることが分かった．今後は，より明瞭性の高い変換目標音声について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 強調技術に導入する．

謝辞 本研究の一部は，JSPS 科研費 26280060 および 24300073 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] 鶴田 他，音講論（秋），pp. 253-254， 2014.
[7] T. Toda *et al., Proc. INTERSPEECH*, pp.632-635, BRIGHTON, UK, 2009.
[8] K. Tanaka *et al., IEICE Trans*, vol. E97-D, no. 6, pp. 1429-1437, 2014.
[9] 近藤 他，信学技報，WIT2007-62，pp. 43-48，2008.
[10] H. Kawahara *et al., Speech Commun.*, vol. 27，no. 3-4，pp. 187-207, 1999.