空気／体内伝導マイクを併用した雑音環境下における非可聴つぶやき強調*

☆田尻祐介, 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

サイレント音声コミュニケーションの実現に向けて，非可聴つぶやき（Nonaudible murmur: NAM）を専用の体表密着型マイクで収録する枠組みが提案されている [1]．しかし，体内伝導音声としてのNAMは，その特性上，通常音声と比べて明瞭性および自然性が大きく劣化する．これを解決するため，統計的手法 [2, 3] によりNAMを通常音声およびささやき声へと変換するNAM強調法が提案されている [4]．ただし，従来の研究では，NAMを遮音室のような静音環境下で収録しているため，実環境へ適用するには，外部雑音の影響を考慮する必要がある．

本稿では，外部雑音に対する頑健性向上に向けて，空気伝導マイクを併用した2チャネル変換法と雑音重畳学習に基づく雑音依存変換モデルの有効性を示す．

2 統計的手法に基づくNAM強調法 [4]

NAMの音響特徴量を通常音声やささやき声といった目標音声の音響特徴量へと変換し，NAMの自然性や明瞭性を改善する．本手法は，学習処理と変換処理から構成される．学習処理では，同一文発声のNAMと目標音声を用い，入出力特徴量間の対応関係を混合正規分布モデル（Gaussian mixture model: GMM）によりモデル化する．入力特徴量はNAMのメルケプストラムセグメント特徴量とする．出力特徴量には，スペクトル特徴量として目標音声のメルケプストラム，音源特徴量として F_0 および非周期成分を用い，それぞれ個別にGMMを学習する．変換処理では，最尤系列変換法 [3] により，NAMの特徴量系列を目標音声の特徴量系列へと変換し，強調音声を得る．

3 外部雑音に頑健なNAM強調法

3.1 空気伝導マイクを併用した2チャネル変換法

NAMは周囲が聴取困難なほど微弱なささやき声であるものの，空気伝導マイクを口唇付近に設置することで，空気伝導収録が可能である．体表密着型マイクで得られる体内伝導NAMと比較し，空気伝導マイクで得られる空気伝導NAMは，体内伝導収録に伴う高域周波数成分の減衰や，スペクトル包絡特性の変化が生じない．そのため，目標音声のスペクトル包絡特性と類似した特徴を持つ信号として利用することを考える．しかし，空気伝導マイクで得られる信号は，周囲の雑音の影響を受けやすい．そこで，体内伝導NAMと空気伝導NAMの両者を入力として用い，雑音に対する脆弱性を回避しつつ，特徴量推定精度の向上を図る．

学習処理では体内伝導NAMと目標音声の間で動的計画法に基づく時間軸伸縮を行い，入出力フレームの対応付けを行う．空気伝導NAMについては，体内伝導NAMに合わせて伸縮させる．体内伝導NAMおよび空気伝導NAMのセグメント特徴量をそれぞれ $\boldsymbol{X}_t^{(\mathrm{b})}, \boldsymbol{X}_t^{(\mathrm{a})}$，目標音声の静的・動的特徴量を $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ とする．対応付けられた各特徴量を用いて，結合確率密度 $P(\boldsymbol{X}_t^{(\mathrm{b})}, \boldsymbol{X}_t^{(\mathrm{a})}, \boldsymbol{Y}_t|\boldsymbol{\lambda})$ をGMMによりモデル化する．ここで，$\boldsymbol{\lambda}$ は各分布の混合重み，平均ベクトルおよび共分散行列から構成されるパラメータセットである．

変換処理では，入力特徴量系列 $\boldsymbol{X}^{(\mathrm{b})}, \boldsymbol{X}^{(\mathrm{a})}$ に対して，条件付き確率密度関数 $P(\boldsymbol{Y}|\boldsymbol{X}^{(\mathrm{b})}, \boldsymbol{X}^{(\mathrm{a})}, \boldsymbol{\lambda})$ を最大にする出力静的特徴量系列 $\hat{\boldsymbol{y}}$ を求める．また，特徴量系列内の変動成分（Global variance: GV）$\boldsymbol{v}(\boldsymbol{y})$ の確率密度関数 $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(\boldsymbol{v})})$ を同時に最大化することで，統計的手法特有の過剰な平滑化処理による影響を低減する．

3.2 雑音依存変換モデル

体表密着型マイクには，空気伝導マイクほどではないものの，外部雑音が混入する．NAMは非常にパワーの小さな音声であるため，空気伝導NAMのみでなく，体内伝導NAMも外部雑音の影響を受ける．結果，静穏環境下で収録されたNAMを用いてGMMを学習した場合，雑音環境下では音響的ミスマッチが生じるため，変換性能が大幅に劣化する．そのため，GMMを雑音環境に適応させる必要がある．

本稿では，GMMを適応する効果の上限を検証するため，雑音依存GMMを構築し，その変換性能を評価する．変換時の雑音は既知として，クリーンなNAMに同一雑音を重畳したデータを生成し，GMMの学習データとして用いる．入出力フレームの対応付けについては，雑音の重畳されていない体内伝導NAMと目標音声で学習したアライメント情報を用いる．変換処理では，雑音重畳学習によりモデル化されたGMMを用いることで，雑音を含むNAMの特徴量をクリーンな目標音声の特徴量へと変換する．

4 実験的評価

4.1 実験条件

男性話者1名の体内伝導NAM（BC-NAM）と空気伝導NAM（AC-NAM）を同時収録する．このとき，NAMとは別に2種類の人混み雑音を収録する．雑音のSNRは，それぞれ空気伝導マイクにおいて −0.5 dB，−11.4 dBである．また，同一話者の通常音声（SP），ささやき声（WH）を空気伝導マイクで収録する．収録文はATR音素バランス文Aセット中の50文とし，40文を学習データ，残りの10文を評価データに用いる．サンプリング周波数は16 kHzとする．

体内伝導NAMおよび空気伝導NAMの特徴量には，FFT分析による0〜24次のメルケプストラム係数から得られるセグメント特徴量（前後4フレーム相当）を用いる．通常音声の分析にはSTRAIGHT分析 [5] を用い，ささやき声の分析にはメルケプストラム分析 [6] を用いる．フレームシフトは5 msとする．GMMの混合数は，メルケプストラム推定用に16，F_0

*NAM enhancement using air- and body-conductive microphones in noisy environments. by TAJIRI, Yusuke, TSURUTA, Sakura, TANAKA, Kou, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, NAKAMURA, Satoshi (NAIST)

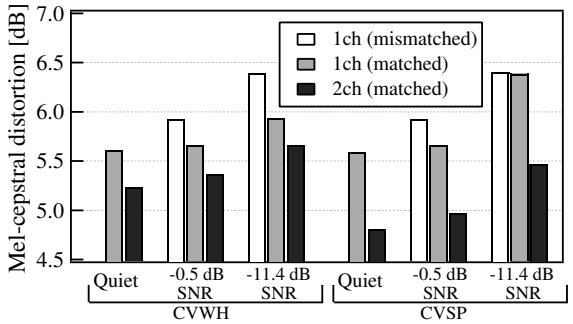


Fig. 1　Mel-cepstral distortion of enhanced speech

推定用に 8，非周期成分推定用に 8 とする．

入力には，体内伝導 NAM および空気伝導 NAM に加え，それらに雑音を重畳して生成したデータを用いる．出力の強調音声は，従来法による変換通常音声（1ch-CVSP）および変換ささやき声（1ch-CVWH），雑音依存モデルを用いた変換通常音声（1ch-CVSP[matched]）および変換ささやき声（1ch-CVWH[matched]），提案法である 2 チャネル変換および雑音依存モデルを用いた変換通常音声（2ch-CVSP[matched]）および変換ささやき声（2ch-CVWH[matched]）の合計 6 種類とする．

4.2　スペクトル特徴量推定精度

環境，手法毎のメルケプストラムひずみを Fig. 1 に示す．なお，メルケプストラムひずみは，0 次項を含まずに計算する．雑音により推定精度が大幅に劣化するが，雑音依存モデルを用いることでその劣化を抑えることができる．また，提案する 2 チャネル変換を用いることで，全ての環境において，1 チャネル変換の推定精度を大幅に改善できることがわかる．

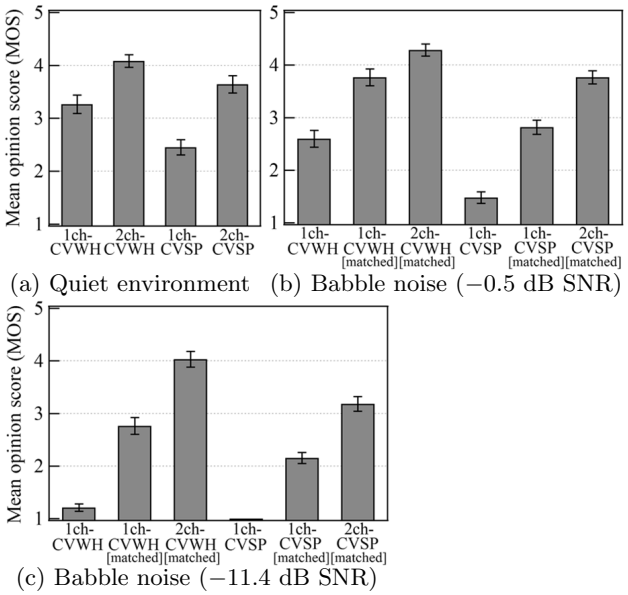
4.3　聞き取りやすさに関する主観評価

強調音声の聞き取りやすさを 5 段階 MOS（1 : 非常に悪い〜5 : 非常に良い）で評価する．被験者は日本人男性 6 名で，1 名あたり手法毎に 22 サンプル，合計 132 サンプル（静穏環境の場合のみ 88 サンプル）を環境ごとに独立に受聴する．

結果を Fig. 2 に示す．雑音依存変換モデルの利用により，強調音声の聞き取りやすさが改善されることがわかる．さらに，すべての環境において，1 チャネル変換よりも 2 チャネル変換による強調音声のほうが聞き取りやすいことがわかる．

次に，強調音声の中で最もスコアの高い 2ch-CVWH と強調前の BC-NAM，AC-NAM に対して，5 段階 MOS 評価を実施する．本実験では，環境による変化についても検証するため，1 名あたり環境，手法毎に 25 サンプル，合計 225 サンプルをまとめて評価する．

結果を Fig. 3 に示す．AC-NAM は，静穏環境下においてはパワーの小さなささやき声のようなものであり，スコアは 4.8 と非常に高いが，雑音の影響を受けやすく，SNR の低下に伴いスコアが大きく減少する．一方，BC-NAM のスコアは静穏環境下で 3.1 と低いものの，−0.5 dB の SNR においてもスコアがほとんど減少しない．このことから，体内伝導収録は空気伝導収録と比べて，外部雑音に頑健であることがわかる．提案法による強調音声（2ch-CVWH）は，雑音環境下において，BC-NAM および AC-NAM のスコアを大幅に上回っており，静穏環境下においても，AC-NAM に若干劣るものの高いスコアが得られている．以上の結果から，提案法の有効性を確認できる．



(a) Quiet environment　(b) Babble noise (−0.5 dB SNR)

(c) Babble noise (−11.4 dB SNR)

Fig. 2　Listenability of enhanced speech (confidence interval 95%)

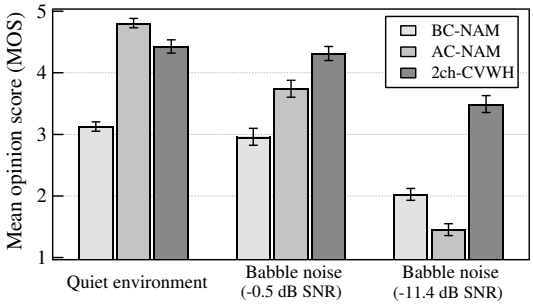


Fig. 3　Listenability of input and enhanced speech (confidence interval 95%)

5　おわりに

外部雑音に頑健な NAM 強調法として，空気伝導マイクの併用による 2 チャネル変換法と雑音依存変換モデルの利用を提案した．実験的評価結果から，静穏環境下および雑音環境下において，スペクトル推定精度が大幅に改善され，NAM の聞き取りやすさを大幅に改善できることを示した．今後は，より実用的な NAM 強調法の実現に向けて，未知の雑音に対する変換モデルの適応や，ロンバード効果が学習・変換処理に及ぼす影響の調査を行う．

謝辞　本研究の一部は，JSPS 科研費 26280060 および 23240023 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.
[2] Y. Stylianou *et al.*, *IEEE Trans.*SAP, Vol 6, No. 2, pp. 131-142, 1998.
[3] T. Toda *et al.*, *IEEE Trans.*ASLP, Vol. 15, No. 8, pp. 2222-2235, 2007.
[4] T. Toda *et al.*, *IEEE Trans.*ASLP, Vol. 20, No. 9, pp. 2505-2517, 2012.
[5] H. Kawahara *et al.*, *Speech Commun.*, Vol. 27, No. 3-4, pp. 187-207, 1999.
[6] K. Tokuda *et al.*, *IEEE Trans.*SAP, Vol. 3, No. 6, pp. 481-489, 1995.