社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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法とその評価

田尻　祐介† 　田中　　宏† 　戸田　智基† 　グラム・ニュービッグ† 　サクリアニ・サクティ†
中村　　哲†

† 奈良先端科学技術大学院大学 情報科学研究科　〒 630–0192 奈良県生駒市高山町 8916–5
E-mail: †{tajiri.yusuke.tk0,ko-t,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp

あらまし　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（Non-Audible 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．しかし，体内伝導収録された音声の音響特徴量は，通常の空気伝導収録された音声のものと異なり，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，統計的手法に基づき，NAM を通常音声やささやき声へと変換する NAM 強調法が提案されている．ただし，従来の研究では，NAM を遮音室のような静穏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．本研究では，外部雑音に対する頑健性向上に向けて，体内伝導マイクおよび通常の空気伝導マイクを併用した 2 チャネル NAM 強調法を提案する．実験的評価結果から，空気／体内伝導マイクを併用し，さらに，外部雑音の混入による直接的な影響と，外部雑音によって引き起こされる発話様式変化とを考慮した変換モデルを構築することで，雑音環境下における音響特徴量変換精度を大幅に改善できることを示す．

キーワード　サイレント音声コミュニケーション，非可聴つぶやき，統計的声質変換，空気／体内伝導音声信号，ロンバード効果

Non-Audible Murmur Enhancement Method using Air- and Body-Conductive Microphones in Noisy Environments and its Evaluation

Yusuke TAJIRI†, Kou TANAKA†, Tomoki TODA†, Graham NEUBIG†, Sakriani SAKTI†, and Satoshi NAKAMURA†

† Graduate School of Information Science, Nara Institute of Science and Technology,
Takayama-cho 8916–5, Ikoma-shi, Nara, 630–0192 Japan
E-mail: †{tajiri.yusuke.tk0,ko-t,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp

Abstract　As one of the silent speech interfaces, Non-Audible Murmur (NAM) microphone which can detect an extremely soft whispered voice has been developed. Although NAM is a promising medium for silent speech communication, its intelligibility and naturalness are significantly degraded by acoustic changes caused by body-conductive recording. To address this issue, several enhancement methods based on statistical voice conversion techniques have been proposed, and their effectiveness has been confirmed in quiet environments. However, it can be expected that NAM will be used not only in quiet, but also in noisy environments, and it is thus necessary to develop enhancement methods that will also work in these cases. In this report, we propose a framework for NAM enhancement using the NAM microphone and an air-conductive microphone. Experimental results demonstrate that the proposed framework is capable of significantly improving enhancement performance in noisy environments by considering not only the effect of noise contamination but also speaking style changes caused by the noise.

Key words　silent speech communication, Non-Audible Murmur, statistical voice conversion, air- and body-conducted speech signal, Lombard effect

1. まえがき

音声コミュニケーションは，我々の生活における重要なコミュニケーション手段の一つであり，携帯電話などの通信端末の普及により，異なる環境下にいる人物との通話も容易になった．しかしながら，聞き手が周囲にいない状況での発声行為は，周囲にとって迷惑になりやすく，また，会話内容が聞こえてしまうという理由から，発声を躊躇するような状況が数多く存在する．これらは，聞き手が聴取可能な声を発さなければならないという制約に起因するものであり，音声コミュニケーションにおける本質的な問題と言える．この問題に対し，聴取可能な声を出さずに音声コミュニケーションを行う技術として，サイレント音声インタフェースが注目を浴びている [1]．

サイレント音声インタフェースの一つとして，非可聴つぶやき（Non-Audible Murmur: NAM）と呼ばれる周囲が聴取困難なほど微弱なささやき声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 [2]．本枠組みは，調音器官のセンシングなどに基づく他の枠組みと比較して，入力に用いられる収録信号が通常音声と類似した特徴を持つため，音声情報の抽出精度が高く，ユーザビリティやコスト面にも優れるという利点を持つ．しかしながら，体内伝導によるローパス特性や口唇の放射特性の欠如などの影響により [3]，体内伝導収録された音声の音響特徴量は通常の空気伝導収録された音声のものと大きく異なる．その結果，体内伝導収録された NAM の明瞭性および自然性は著しく劣化するため，そのまま通話に利用するのは困難である．

NAM の明瞭性および自然性を改善する技術として，統計的声質変換 [4], [5] に基づき，NAM を通常音声やささやき声へと変換する NAM 強調法が提案されている [6]．本手法では，同一文発声の NAM と目標音声のパラレルデータから，入出力特徴量間の対応関係を事前に学習しておくことで，任意の発話内容に対する変換を可能とする．従来の研究では，遮音室のような静穏環境下で収録された NAM に対して，NAM 強調法の有効性が示されている．一方で，実環境へ適用するには，実環境下で起こり得る要因を考慮する必要がある．特に，外部雑音は実環境において必ず存在し，収録信号に混入した場合，従来の強調法に対して著しい性能劣化を引き起こす．

本研究では，外部雑音に対する頑健性向上に向けて，体内伝導マイクおよび空気伝導マイクで収録した NAM を統計的声質変換の入力として用いる 2 チャネル NAM 強調法を提案する．さらに，提案法において，変換モデルを雑音環境に適応させた際に得られる効果の上限を検証するため，雑音依存変換モデルを構築する．また，ロンバード効果 [7], [8] として知られるような外部雑音に伴う発話様式の変化にも対処するために，発話様式変化の影響を考慮した変換モデルを構築する．実験的評価では，構築した各変換モデルに対して，実際の雑音環境下での発話を想定して収録した NAM を入力し，その変換精度を比較する．結果として，提案法である 2 チャネル NAM 強調において，混入外部雑音および発話様式変化の両者を考慮した変換モデルを構築することで，大幅な性能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 [6]

体内伝導収録された NAM の音響特徴量を通常音声やささやき声といった自然な音声の音響特徴量へと変換することにより NAM の明瞭性や自然性を改善する．本手法は，以下に示す学習処理と変換処理から構成される．

2.1 学習処理

各時間フレームにおける入出力の静的特徴量ベクトルをそれぞれ $\boldsymbol{x}_t$, $\boldsymbol{y}_t$ とする．ここで，t はフレーム番号を表す．入力特徴量には，体内伝導収録に伴う音素情報の抜け落ちなどを補償するため，当該フレームおよび前後 $\pm L$ フレームを用いて計算されたセグメント特徴量ベクトル $\boldsymbol{X}_t = \boldsymbol{A}[\boldsymbol{x}_{t-L}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+L}^\top]^\top + \boldsymbol{b}$ を用いる．ここで，$\top$ は転置を表す．$\boldsymbol{A}$ および $\boldsymbol{b}$ はそれぞれ主成分分析により得られる固有ベクトル，バイアスベクトルである．出力特徴量には，結合静的・動的特徴量ベクトル $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ を用いる．

学習処理では，入力音声と出力音声の同一発話対で構成されるパラレルデータに対して，フレーム間の対応付けを行うことで得られる結合入力・出力特徴量ベクトル $[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ を次式に示す混合正規分布モデル（Gaussian mixture model: GMM）でモデル化する．

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}) = \sum_{m=1}^{M} w_m \mathcal{N}([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}) \tag{1}$$

ここで，$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均ベクトル $\boldsymbol{\mu}$ および共分散行列 $\boldsymbol{\Sigma}$ の正規分布を示す．また，m は分布番号，M は分布数を表す．GMM のパラメータセットは $\boldsymbol{\lambda}$ で示され，各分布に対する重み w_m，平均ベクトル $\boldsymbol{\mu}_m^{(X,Y)}$ および全共分散行列 $\boldsymbol{\Sigma}_m^{(X,Y)}$ で構成される．さらに，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ため，出力静的特徴量ベクトルの系列内変動（Global variance: GV）$\boldsymbol{v}(\boldsymbol{y})$ の確率密度関数を，次式に示す正規分布でモデル化する．

$$P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}) = \mathcal{N}(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(v)}) \tag{2}$$

パラメータセット $\boldsymbol{\lambda}^{(v)}$ は平均ベクトル $\boldsymbol{\mu}^{(v)}$ および対角共分散行列 $\boldsymbol{\Sigma}^{(v)}$ から構成される．

2.2 変換処理

変換処理では，最尤系列変換法 [5] 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する．まず，入力特徴量系列 $\boldsymbol{X}$ に対して，尤度 $P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda})$ を最大にする分布系列 $\hat{\boldsymbol{m}}$ を決定する．次に，次式に示すような GV を考慮した目的関数を最大にする出力静的特徴量系列を求める．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}) P(\boldsymbol{v}(\boldsymbol{y})|\lambda^{(v)})^\omega \quad \text{subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{3}$$

ここで，$\boldsymbol{W}$ は静的特徴量系列を静的・動的特徴量系列へと拡張する変換行列，ω は尤度重みを表す．

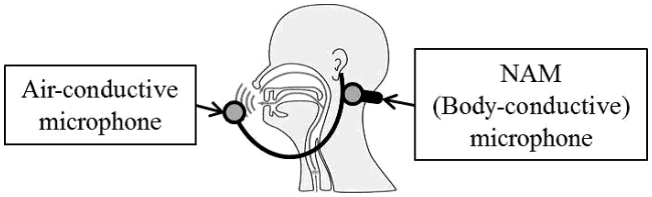

図 1　空気伝導マイクおよび体内伝導マイクの配置図

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 NAM 強調法

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，図 1 に示すように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．体内伝導 NAM と比較し，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が可能である．しかしながら，空気伝導収録は外部雑音の影響を受けやすいのに対し，体内伝導収録はマイクの構造上，外部雑音に対して比較的頑健である．そこで，このように相補的な特徴を持つ体内伝導 NAM と空気伝導 NAM の両者を統計的声質変換の入力として用いることで，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．

まず，空気伝導 NAM および体内伝導 NAM それぞれについて，**2.1** と同様の手順で，セグメント特徴量ベクトルを計算する．次に，求めた各セグメント特徴量ベクトル $\boldsymbol{X}_t^{(\mathrm{a})}$，$\boldsymbol{X}_t^{(\mathrm{b})}$ を一つの特徴量ベクトル $\boldsymbol{X}_t^{(\mathrm{a,b})} = [\boldsymbol{X}_t^{(\mathrm{a})\top}, \boldsymbol{X}_t^{(\mathrm{b})\top}]^\top$ に結合し，入力特徴量として用いる．残りの処理については従来法と同様の手順で行う．

3.2 雑音依存変換モデル

体内伝導マイクには，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差異が生じるため，変換性能が大幅に劣化する．この影響を軽減するため，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．図 2 に学習処理の流れを示す．まず，NAM と雑音を別々に収録する．次に，変換時の雑音は既知として，クリーンな NAM に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，雑音を重畳していないクリーンな体内伝導 NAM と目標音声で学習したアライメント情報を用いる．

3.3 ロンバード効果を考慮した変換モデル

雑音環境下で音声を収録する場合，雑音の音量が大きくなるにつれて，発話者が自身の発話内容を聴取することが困難となる．このとき，十分な聴覚フィードバックを得るために，無意識のうちに発話様式を変化させてしまう現象がロンバード効果

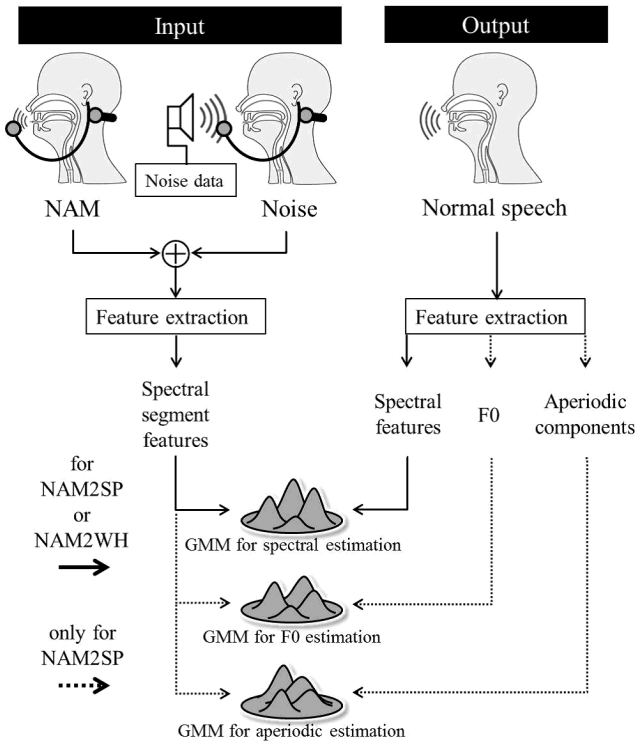

図 2　雑音依存 GMM の学習処理（NAM2SP は NAM から通常音声への変換，NAM2WH は NAM からささやき声への変換を表す）

として知られており，周波数特性やフォルマント，発話の音量，音素の継続時間など様々な音響的特徴が変化する．NAM は非常にパワーの小さな音声であるため，雑音環境下で収録する際は，このロンバード効果の影響を強く受ける [9]．そのため，統計的手法に基づく NAM 強調法を雑音環境へ適用するには，収録信号への外部雑音混入だけでなく，ロンバード効果の影響も考慮する必要がある．

本稿では，**3.1** の雑音依存 GMM に加えて，ロンバード効果を考慮した雑音依存 GMM を構築し，その変換性能の評価および比較を行う．図 3 に学習処理の流れを示す．まず，雑音依存 GMM の学習と同様に，NAM と雑音を別々に収録する．このとき，ヘッドフォンを用いて，雑音収録時と同じ音量になるよう調整した雑音を発話者に提示しながら NAM を収録することで，ロンバード効果を模擬的に再現する．次に，ロンバード効果の影響を受けたクリーンな NAM に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．

4. 実験的評価

4.1 実 験 条 件

男性話者 1 名の NAM を空気伝導マイクおよび体内伝導マイクで同時収録する．また，同一話者の通常音声，ささやき声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 [10]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．雑音依存モデルを構築するため，次の 7 種類の雑音を NAM とは別に各マイクで同時収録する．

- 50 dB のバブル雑音（babble50dB）
- 60 dB のバブル雑音（babble60dB）

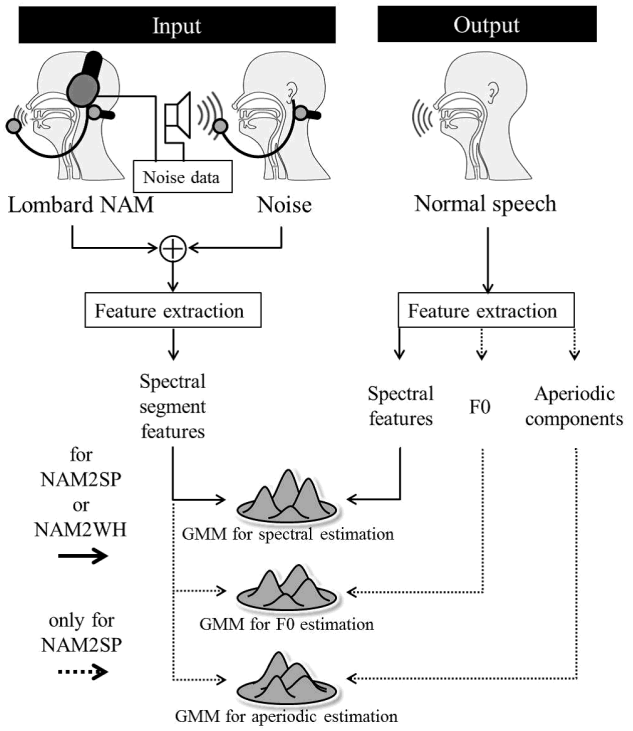

図 3　ロンバード効果を考慮した雑音依存 GMM の学習処理（NAM2SP は NAM から通常音声への変換，NAM2WH は NAM からささやき声への変換を表す）

- 70 dB のバブル雑音（babble70dB）
- 50 dB のオフィスの空調雑音（office50dB）
- 60 dB の人混み雑音 （crowd60dB）
- 70 dB の展示場の雑音 （booth70dB）
- 80 dB の地下鉄構内の雑音 （station80dB）

雑音の音量は発話者の頭部位置で測定した値である．なお，バブル雑音は，男女それぞれ 10 名の音声を重畳して生成したヒューマンスピーチライク雑音 [11] とする．ロンバード効果の再現では，音量のみが異なる 3 種類のバブル雑音をヘッドフォンで発話者に提示しながら，NAM を各マイクで同時収録する．また，実際の雑音環境下での発話を想定した収録では，上述した 7 種類の雑音をスピーカーで発話者に提示しながら，NAM を各マイクで同時収録する．収録信号のサンプリング周波数は全て 16 kHz とする．

スペクトル特徴量には，0〜24 次のメルケプストラム係数を用い，空気／体内伝導 NAM，通常音声，ささやき声，それぞれに対して，FFT 分析，STRAIGHT 分析 [12]，メルケプストラム分析 [13] を用いる．セグメント特徴量は 50 次元とし，当該フレームおよび前後 4 フレームを用いて計算する．音源特徴量には，STRAIGHT 分析 [14], [15] で抽出した対数 F_0 および各周波数帯域（0〜1，1〜2，2〜4，4〜6，6〜8 kHz）の非周期成分を用いる [16]．分析フレームシフト長は 5 ms とする．NAM から通常音声への変換（NAM2SP）では，スペクトル変換用，F_0 変換用，非周期成分変換用として計 3 つの GMM を学習し，その混合数はそれぞれ 32，16，16 とする．NAM からささやき声への変換（NAM2WH）では，スペクトル変換用として 1 つの GMM を学習し，その混合数は 32 とする．

4.2　混入外部雑音のみを考慮した変換モデルの評価

3.2 の図 2 の手順で構築した変換モデルに対して，1）クリーンな NAM と雑音を重畳して生成した信号，2）実際の雑音環境下での発話を想定して収録した NAM，を入力し，両者の変換精度を比較する．変換精度の評価尺度として，スペクトル特徴量に対してはメルケプストラムひずみ，音源特徴量に対しては F_0 の有声無声判別誤り率および相関係数，非周期成分ひずみを用いる．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．スペクトル特徴量の変換精度を図 4 および図 5 に，音源特徴量の変換精度を表 1 および表 2 に示す．なお，表中の太字は，1 チャネル変換と 2 チャネル変換の結果を比較したときの，より精度が高い方を表す．

まず，1）の NAM と雑音を重畳して生成した信号を入力した場合，図 4 より，従来法による変換（1ch mismatched）と比較して，混入外部雑音を考慮した変換（1ch matched）では，雑音の種類や変換対象に関係なく，メルケプストラムひずみが減少していることがわかる．また，2 チャネル変換（2ch matched）を用いることで，変換精度がさらに改善されている．特に，通常音声への変換では，70 dB のバブル雑音（babble70dB）や展示場の雑音（booth70dB）に対する結果において，静穏環境下で従来法を用いた際の結果を大幅に上回るスペクトル変換精度が得られている．音源特徴量についても，表 1 に示す通り，スペクトル特徴量ほどではないものの 2 チャネル変換による改善がみられ，60 dB のバブル雑音（babble60dB）での有声無声判別を除く全ての結果において，2 チャネル変換が 1 チャネル変換を上回っている．

次に，2）の実際の雑音環境下での発話を想定して収録した NAM を入力した場合，図 5 の結果では，図 4 のものと異なる傾向がみられる．図 4 では，2 チャネル変換による改善度合いが雑音の種類に関係なくほぼ一定であるのに対し，図 5 では，雑音の音量が増加するにつれて，2 チャネル変換での変換精度が著しく低下している．特に，80 dB の地下鉄構内の雑音（station80dB）では，変換対象に関係なく，2 チャネル変換と 1 チャネル変換の結果が逆転する．音源特徴量についても同様の傾向がみられ，表 2 より，雑音の音量の増加に伴い，全体的に変換精度が低下するのがわかる．また，70 dB の雑音に対しては，有声無声判別および非周期成分ひずみで 2 チャネル変換と 1 チャネル変換の結果が逆転する．

これらの結果から，2 チャネル変換と混入外部雑音を考慮した変換モデルは有効であることがわかる．一方で，NAM と雑音を重畳する場合と雑音環境下で NAM を収録する場合の結果の比較から，NAM と雑音を重畳する方法では，十分な精度で実際の雑音環境下での収録信号を模擬できないことがわかる．

4.3　外部雑音混入およびロンバード効果を考慮した変換モデルの評価

3.3 の図 3 の手順で構築した変換モデルに対して，1）クリーンなロンバード NAM と雑音を重畳して生成した信号，2）実際の雑音環境下での発話を想定して収録した NAM，を入力し，両者の変換精度を比較する．スペクトル特徴量の変換精度を図

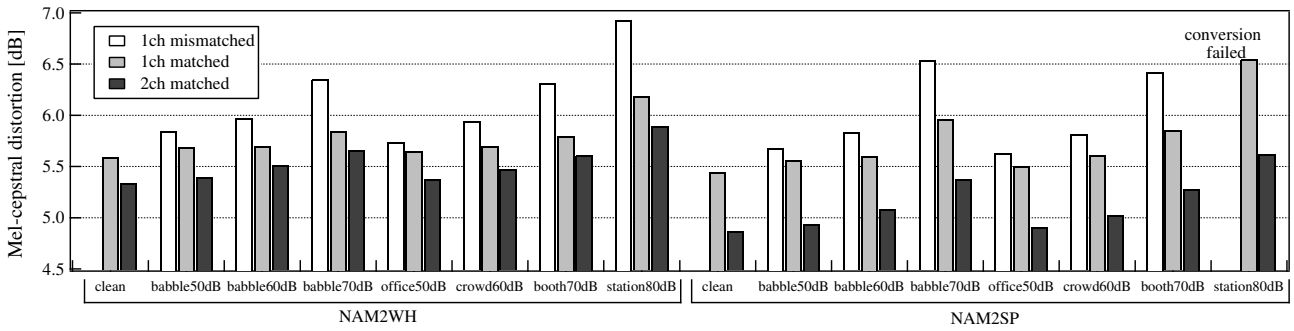

図 4　スペクトル特徴量の変換精度（NAM と雑音を重畳して生成した信号を入力した場合）

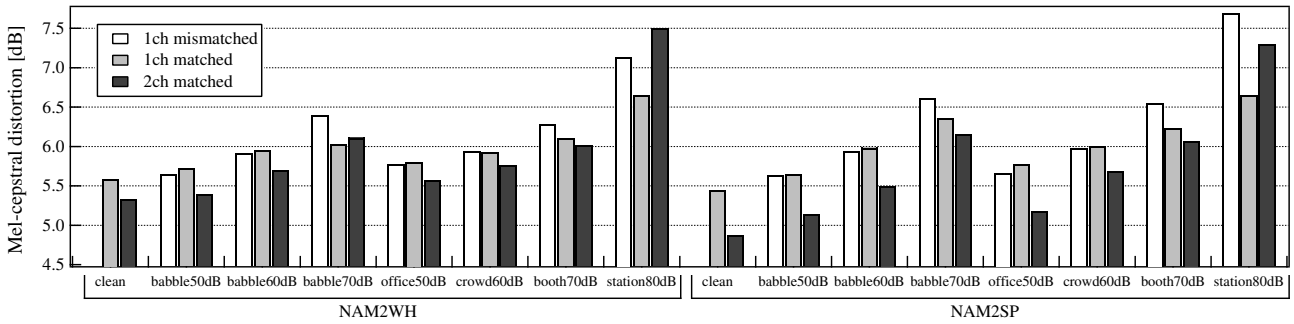

図 5　スペクトル特徴量の変換精度（雑音環境を想定して収録した NAM を入力した場合）

6 および図 7 に，音源特徴量の変換精度を表 3 および表 4 に示す．

まず，1）のロンバード NAM と雑音を重畳して生成した信号を入力した場合，図 6 より，雑音の種類や変換対象に関係なく，2 チャネル変換によってメルケプストラムひずみが減少しているのがわかる．また，図 4 では，雑音の音量の増加に伴い，ひずみも増加しているのに対し，図 6 では，雑音の音量に依存せず，ほぼ一定になっている．これは，周囲の雑音の音量が大きくなると，発話の音量も大きくなり，SN 比がある程度一定に保たれるためだと考えられる．音源特徴量については，表 3 より，有声無声判別以外の結果において，2 チャネル変換による改善がみられる．雑音の音量による違いについては，スペクトル特徴量のように一定の値とはならないものの，表 1 の結果のような音量の増加に伴う変換精度の低下はみられない．

次に，2）の実際の雑音環境下での発話を想定して収録した NAM を入力した場合，図 7 と図 5 と比較すると，変換精度が全体的に改善していることがわかる．これは，ロンバード効果を考慮したことによる結果であり，ロンバード効果の影響が強くあらわれる 70 dB の雑音（babble70dB）に対する結果では，ささやき声への 2 チャネル変換で，約 6 dB から 5.7 dB 以下に，通常音声への 2 チャネル変換で，約 6 dB から 5.4 dB 以下にひずみが減少している．一方で，ロンバード効果の影響があらわれにくい 50 dB の雑音（babble50dB）に対する結果では，変換精度に大きな変化がみられない．音源特徴量については，表 4 と表 2 を比較すると，ロンバード効果を考慮したことにより，F_0 の相関値を除いて全体的に変換精度が改善されているのがわかる．ただし，これまでの結果でみられたような 2 チャネル変換の優位性は確認できない．

表 1　音源特徴量の変換精度（NAM と雑音を重畳して生成した信号を入力した場合）

	U/V error [%]		F_0 correlation		Aperiodic dist. [dB]	
	1ch	2ch	1ch	2ch	1ch	2ch
clean	24.2	**20.8**	0.33	**0.34**	4.95	**4.69**
babble50dB	25.5	**19.5**	0.30	**0.35**	5.07	**4.72**
babble60dB	**23.0**	24.6	0.29	**0.31**	4.93	4.93
babble70dB	32.9	**28.7**	0.21	**0.26**	5.14	**5.08**

表 2　音源特徴量の変換精度（雑音環境を想定して収録した NAM を入力した場合）

	U/V error [%]		F_0 correlation		Aperiodic dist. [dB]	
	1ch	2ch	1ch	2ch	1ch	2ch
clean	24.2	**20.8**	0.33	**0.34**	4.95	**4.69**
babble50dB	28.3	**19.6**	0.30	**0.31**	5.27	**4.94**
babble60dB	29.5	**27.7**	0.20	**0.25**	5.60	**5.39**
babble70dB	**34.9**	44.7	0.23	**0.28**	**5.74**	6.03

これらの結果から，混入外部雑音およびロンバード効果の両者を考慮した変換モデルは有効であることがわかる．また，ヘッドフォンで発話者に雑音を提示しながら収録したロンバード NAM に対して雑音を重畳する方法により，実際の雑音環境下での収録信号をより高い精度で模擬できることがわかる．一方で，雑音の音量の増加に伴い，変換精度の低下が生じることから，模擬した収録信号と実際の収録信号の違いは大きくなる傾向がみられる．このことから，混入外部雑音やロンバード効果以外にも考慮すべき要因が残されていることがわかる．

5. おわりに

外部雑音に頑健な NAM 強調法の実現に向けて，空気伝導マ

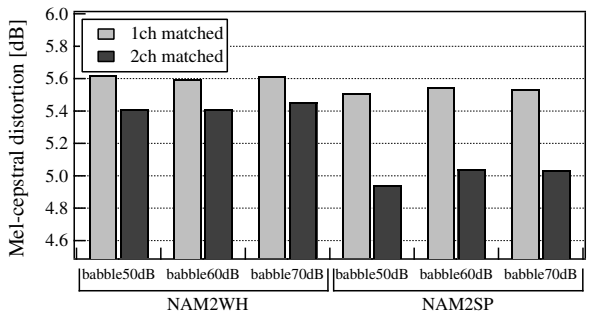

図 6　ロンバード効果を考慮した場合のスペクトル特徴量の変換精度（NAM と雑音を重畳して生成した信号を入力した場合）

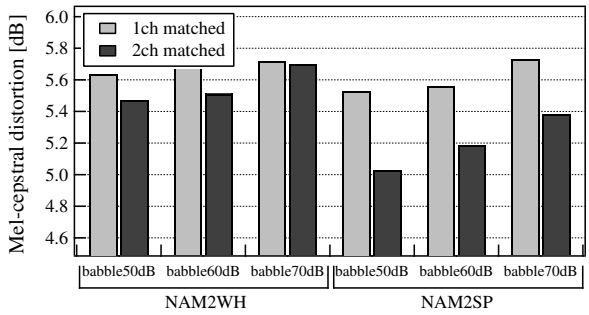

図 7　ロンバード効果を考慮した場合のスペクトル特徴量の変換精度（雑音環境を想定して収録した NAM を入力した場合）

表 3　ロンバード効果を考慮した場合の音源特徴量の変換精度（NAM と雑音を重畳して生成した信号を入力した場合）

	U/V error [%]		F_0 correlation		Aperiodic dist. [dB]	
	1ch	2ch	1ch	2ch	1ch	2ch
babble50dB	22.3	**20.3**	0.29	**0.30**	5.04	**4.81**
babble60dB	**24.3**	26.0	**0.26**	0.25	5.04	**5.00**
babble70dB	**23.3**	24.9	0.22	**0.29**	5.01	**4.97**

表 4　ロンバード効果を考慮した場合の音源特徴量の変換精度（雑音環境を想定して収録した NAM を入力した場合）

	U/V error [%]		F_0 correlation		Aperiodic dist. [dB]	
	1ch	2ch	1ch	2ch	1ch	2ch
babble50dB	23.7	**21.2**	0.19	**0.32**	5.10	**4.77**
babble60dB	**22.8**	27.5	0.20	**0.23**	**4.96**	5.13
babble70dB	**29.9**	31.5	**0.24**	0.15	5.31	**5.26**

イクおよび体内伝導マイクで収録した NAM を，統計的声質変換の入力として用いる 2 チャネル強調法を提案した．外部雑音の混入だけでなく，雑音環境下で引き起こされる発話様式変化も考慮し，変換モデルを構築することで，雑音環境下でも静穏環境下で従来法を使用した場合と同等のスペクトル変換精度が得られることを示した．また，雑音環境を想定して行った実験の結果から，外部からの雑音混入や発話様式変化以外にも考慮すべき要因が存在することが示唆されており，さらなる調査が必要であることがわかった．今後は，それらの調査や未知の雑音に対する変換モデルの適応技術の構築に取り組む．

謝　辞

本研究の一部は，JSPS 科研費 15K12064 および 26280060 の助成を受け実施したものである．

文　献

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, “Silent speech interfaces,” *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[2] Y. Nakajima, H. Kashioka, N. Cambell, and K. Shikano, “Non-Audible Murmur (NAM) recognition,” *IEICE Transactions on Information and Systems*, vol. E89-D, no. 1, pp. 1–8, 2006.

[3] T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima, and K. Shikano, “Silent-speech enhancement using body-conducted vocal-tract resonance signals,” *Speech Communication*, vol. 52, no. 4, pp. 301–313, 2010.

[4] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[5] T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[6] T. Toda, M. Nakagiri, and K. Shikano, “Statistical voice conversion techniques for body-conducted unvoiced speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, 2012.

[7] J. C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers,” *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[8] J. C. Junqua, “The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex,” *Speech Communication*, vol. 20, no. 1, pp. 13–22, 1996.

[9] T. Toda, K. Nakamura, T. Nagai, T. Kaino, Y. Nakajima, and K. Shikano, “Technologies for processing body-conducted speech detected with Non-Audible Murmur microphone,” *Proc. INTERSPEECH*, pp. 632–635, 2009.

[10] Y. Sagisaka, K. Takeda, M. Ave, S. Katagiri, T. Umeda, and H. Kuwabara, “A large-scale Japanese speech database,” *First International Conference on Spoken Language Processing*, pp. 1089–1092, 1990.

[11] 梶田将司, 小林大祐, 武田一哉, 板倉文忠, “ヒューマンスピーチライク雑音に含まれる音声的特徴の分析,” 日本音響学会誌, vol. 53, no. 5, pp. 337–345, 1997.

[12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[13] 徳田恵一, 小林隆夫, 深田俊明, 斎藤博徳, 今井聖, “メルケ プストラムをパラメータとする音声のスペクトル推定,” 信学論 (A), vol. J74-A, no.8, pp. 1240–1248, 1991.

[14] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity,” *Proc. EUROSPEECH*, vol. 99, no. 6, pp. 2781–2784, 1999.

[15] H. Kawahara, J. Estill, and O. Fujimaru, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT,” *Proc. MAVEBA*, pp. 59–64, 2001.

[16] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation,” *Proc. INTERSPEECH*, pp. 2266–2269, 2006.