

話者適応に基づく複数話者の非可聴つぶやき認識 における話者正規化学習の有効性

長井 孝之[†], 中村 圭吾[†], 戸田 智基[†], 猿渡 洋[†], 鹿野 清宏[†]

[†] 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: {takayuki-n,kei-naka,tomoki,sawatari,shikano}@is.naist.jp

あらまし 近年, 周囲の者に知覚されにくい音声認識を実現するための技術として, 非可聴つぶやき (Non-Audible Murmur: NAM) 認識が注目されている. これまで, NAM 発声に熟知した話者に対して NAM 用特定話者音響モデルを作成することで, 高い認識性能が得られている. 一方で, 一般的な話者, すなわち NAM 発声に不慣れな話者の認識性能は, 話者毎に大きくばらつくことが分かっている. 本稿では, 不慣れな話者の認識性能を向上させ, 話者毎の認識性能のばらつきを抑えるために, 各話者への適応時に, 話者正規化学習 (Speaker Adaptive Training: SAT) を用いた正準モデルを使用する. 実験的評価により, SAT を使用することで認識率が大きく改善されることを示す.

キーワード Non-Audible Murmur(NAM), 話者適応, 話者正規化学習, 音響モデル

Effectiveness of Speaker Adaptive Training in Non-Audible Murmur Recognition Based on Speaker Adaptation for Various Speakers

Takayuki NAGAI[†], Keigo NAKAMURA[†], Tomoki TODA[†], Hiroshi SARUWATARI[†], and
Kiyohiro SHIKANO[†]

[†] Graduate School of Information Science, Nara Institute of Science and Technology
Takayama-cho 8916-5, Ikoma-shi, Nara, 630-0192 Japan
E-mail: {takayuki-n,kei-naka,tomoki,sawatari,shikano}@is.naist.jp

Abstract Non-Audible Murmur (NAM) recognition is a potential technique to realize a silent speech interface. In previous work, it has been reported that high recognition rates are achieved by constructing speaker-dependent NAM acoustic models for a few speakers who know how to utter NAM so that it is well recognized. On the other hand, it is known that the variance of recognition rates for general people, who are not familiar with the way to utter NAM, is relatively large in NAM recognition. In this paper, to improve the NAM recognition accuracy for those people and to reduce the variance of the accuracies over different speakers, we refine the initial model of adaptation using speaker adaptive training (SAT). Experimental results demonstrate the effectiveness of introducing SAT in NAM recognition.
Keyword Non-Audible Murmur(NAM), speaker adaptation, speaker adaptive training, acoustic model

1 はじめに

音声は、最も一般的なコミュニケーション手段の1つである。音声を用いて、人間と機械との自然なコミュニケーションを実現するために、音声認識に関する数多くの研究が進められてきた。しかしながら、雑音環境下では、音声認識の性能が急激に劣化するという問題や、オフィスや図書館といった静環境下では、ユーザーが発話行為自体に躊躇したり発声行為が周囲の人々の迷惑になるという根本的な問題は未解決のままである。

音声インターフェースが内包するこうした問題を解決するために、体表から直接信号を採取するデバイスを用いた音声認識の研究が進められている。雑音に対して頑健な音声認識を行うために、Zhengら [1] は、骨伝導マイクロフォンと通常の気導音マイクロフォンを併用する手法を提案している。Jouら [2] は、周囲の者が知覚しにくい音声認識を行うために、Throat microphone を使用して、ささやき声の認識を試みている。さらに、周囲の者に完全に知覚されない音声認識を実現するために、音声以外の信号を用いる手法も提案されている。その中の一つとして、Maier-Heinら [3] は、筋電 (Electromyography: EMG) を用いるシステムを提案している。

近年注目を集めている新しい体表密着型マイクロフォンとして、中島ら [4] は、聴診器の原理を応用した非可聴つぶやき (Non-Audible Murmur: NAM) マイクロフォンを開発している。NAMとは、発話者の近くにいる人でも内容を聴取することが困難なほど小さなつぶやき声を、体表に圧着したマイクロフォンで採取した信号 [4] である。NAMマイクロフォンは、外部雑音に対して頑健であり、NAMを採取することに秀でていることから、周囲の雑音に頑健かつ、周囲の者に知覚されにくい音声認識を実現するために、我々は、このNAMマイクロフォンの応用に注目している。

先行研究では、NAMの発声方法を熟知している特定話者に対して、NAM認識の性能が示されている [4, 5]。一方で、より一般的な話者、すなわちNAMの発声方法に不慣れな話者に対する認識性能は話者毎に大きく異なり、話者によっては話者適応処理を施しても高い認識性能が得られない事が分かっている [6]。

本稿においては、NAM発声に不慣れな話者にお

ける特定話者音響モデルの性能をさらに向上させ、認識性能のばらつきを抑えることを目指す。話者正規化学習 (Speaker Adaptive Training: SAT) [7] によって従来より話者適応に適した初期モデルを作成し、その初期モデルから各話者へ適応する。適応後の特定話者音響モデルを用いて大語彙連続音声認識実験を行い、NAM認識におけるSATの有効性を評価する。

本稿では、2でNAM認識について述べ、3で使用するNAMデータに関して述べる。4でSATを用いて話者適応時の初期モデルを作成する手法を説明し、5でNAM認識におけるSATの有効性を実験的に評価する。最後に6で本稿をまとめる。

2 NAM 認識

2.1 NAM の特徴

NAMの信号は、NAMマイクロフォンを耳介後方部の肌に直接圧着させることで採取される。このようにして得られたNAMは、空気伝導通常音声と比べると以下の点で特徴的である。

1. NAMは声帯振動がないため、取得される信号に基本周波数は含まれない。
2. 体内伝導特性のため、採取された信号には、口唇の放射特性が含まれず、約4kHz以上の高域のパワーの減衰が激しい。
3. NAMの音響特徴量は、NAMの収録環境 (NAMマイクロフォンの装着方法等) に敏感である。
4. NAMは、不慣れな人にとっては特殊な発声なので、発声に対する慣れがある程度必要である。

音声認識においては、スペクトル包絡に含まれる音韻情報が重要であるため、特徴1は、大きな問題ではない。特徴2に関して、気導音と比べるとスペクトルに含まれる情報量が少ないが、音声認識に必要な情報は含まれている。特徴3、及び4はNAMの使用法と話者に依存する特徴であるため、NAM認識を利用する場合、これらの問題点を考慮する必要がある。

2.2 従来研究

従来研究では、混合正規分布を出力確率密度分布とする隠れマルコフモデル (Hidden Markov Model: HMM) に基づく、NAMの大語彙連続音声

認識が行われている。特定話者の数千発話を用いて、NAM用モノフォンモデルを作成し、2万語彙の新聞記事読み上げタスクにおいて、90%以上の単語正解精度が得られている [4]。また、学習用のNAMデータが数百発話と比較的少ない場合でも、予めよく学習された通常音声用不特定話者音響モデルを初期モデルとし、最尤線形回帰 (Maximum Likelihood Linear Regression: MLLR) [8] による適応を繰り返すことで、高精度な特定話者モデルの構築が可能であり、数千発話の学習データを用いるのと同等の性能が得られている [5]。従来行われてきたこれらの研究では、NAM認識の可能性を示すために、NAM発声を熟知した話者のデータを評価している。

我々はこれまでに、NAM発声に不慣れた話者に対するNAM認識の可能性を調査してきた [6]。様々な話者のNAM発声に対する特定話者音響モデルを作成するには、学習データを収集するコストを考慮すると、少量の適応データを用いて各話者へMLLR適応を繰り返す行方手法が効率的であると考えられる。その際の初期モデルとして、通常音声用の不特定話者音響モデルか、適応話者以外のNAMデータも有効利用して通常音声用の不特定話者音響モデルを再学習したモデルを使用した際の性能評価を行った。初期モデルの再学習の手法としては、MLLRと最大事後確率 (Maximum A Posteriori: MAP) 推定 [9] を併用した手法や、Baum-Welchアルゴリズムを用いており、Baum-Welchアルゴリズムを用いて再学習した初期モデルの方がよい性能が得られることが分かっている。しかし、どの初期モデルを用いた場合でも、話者毎の認識性能には依然大きなばらつきが存在する。

3 使用するNAMデータ

本稿で評価の対象とする話者は、NAM発声に不慣れた話者58名である。各話者の音素バランス文約50~60発話と新聞記事読み上げ文約120~170発話のNAMを収録する¹。収録は、収録話者と収録音声を確認する立会人の2名で行う。収録前に、NAMの発声方法やNAMマイクロフォンの装着位置に関して、立会人から収録話者へ指示する。収録中は、立会人が収録音声聞き、収録音声の明瞭性、雑音の混入、言い間違い等を常時確認する。

¹ 発声をしやすくするため、長い読み上げ文は適当な長さに人手で分割し、複数の発話として収録する。

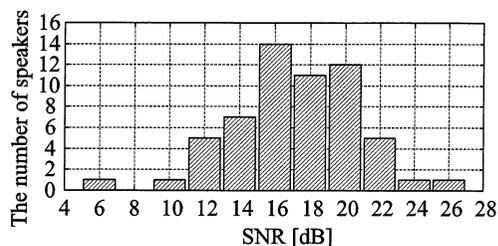


Fig. 1 A histogram of SNR of NAM data uttered by each speaker. Each interval is set to ± 1 dB.

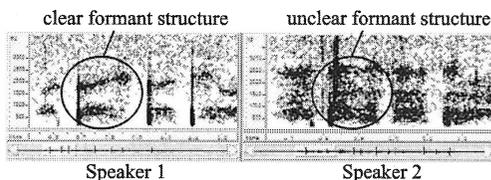


Fig. 2 An example of spectrograms uttered by two different speakers.

何らかの問題がある場合、その発話を再度収録する。NAM収録は全て防音室で行う。このようにして収録したデータを音響モデルの学習データ及び認識時の評価データとして使用する。

Fig. 1は、実際に収録した各話者毎のNAMデータの平均SNRのヒストグラムを示している。収録したデータ内で背景雑音は一定であると仮定し、無音区間の平均パワーを雑音パワーとして、SNRを計算している。Fig. 1を見ると、話者毎にSNRが大きく異なっており、現状では、一貫したSNRでの収録が困難であることが分かる。また、ある2名のNAMのスペクトログラムをFig. 2に示す。発声内容は同一であるにも関わらず、2名の話者のフォルマントの明瞭度は大きく異なっているのが分かる。これらの差はマイクロフォンの装着方法やNAMの発声方法の差に起因すると考えられる。

4 SATに基づく適応時の初期モデル構築手法

本稿では、従来の初期モデルを更に改善し、各話者へ適応後の特定話者音響モデルの性能を向上させるために、初期モデルの作成にNAMデータを用いたSATを導入する。SATの処理の流れは以下の通りである。

1. ある元モデルから各学習話者の特定話者音響

モデルへの写像を推定する.

2. 推定された写像の逆写像を用いて, 各学習話者のデータを変換する.
3. 変換されたデータを用いて元モデルのモデルパラメータを更新し, 正準モデルを作成する.

上述した写像を推定するのに, 制約付最尤線形回帰 (Constrained Maximum Likelihood Linear Regression: CMLLR) [10] を用いることで, 非常に効率的に正準モデルを学習することが可能となる. CMLLR では, ある HMM の状態の m 番目の出力確率密度分布の平均ベクトル μ_m と共分散行列 Σ_m を, 以下のように線形変換する変換行列およびバイアスペクトルが推定される.

$$\hat{\mu}_m^{(i)} = \mathbf{H}^{(i)} \mu_m + \bar{\mathbf{b}}^{(i)} \quad (1)$$

$$\hat{\Sigma}_m^{(i)} = \mathbf{H}^{(i)} \Sigma_m \mathbf{H}^{(i)T} \quad (2)$$

ここで i は話者のインデックスを表す. このとき, EM アルゴリズムにおける補助関数に式 (1), 式 (2) を代入すると, CMLLR によって推定された変換の逆変換を用いて学習データの特徴量を変換して元モデルのモデルパラメータの学習を行う枠組みになる. つまり, $\bar{\mathbf{b}}^{(i)} = -\mathbf{H}^{(i)-1} \hat{\mu}_m^{(i)}$, $\mathbf{A}^{(i)} = \mathbf{H}^{(i)-1}$ とおき, 学習データの特徴量 $\mathbf{o}^{(i)}$, 変換後の特徴量 $\hat{\mathbf{o}}^{(i)}$ とすると,

$$\hat{\mathbf{o}}^{(i)} = \mathbf{A}^{(i)} \mathbf{o}^{(i)} + \bar{\mathbf{b}}^{(i)} \quad (3)$$

のように学習データの特徴量を変換し, 全話者の変換学習データ $\hat{\mathbf{o}}$ を学習に用いる.

5 実験的評価

NAM 認識において, SAT によって作成される初期モデルの有効性を示すために, 各種初期モデルから各話者へ繰り返し MLLR 適応を行い, 特定話者音響モデルを作成する. 作成されたモデルを使用して, NAM の大語彙連続音声認識を行い, 各話者の認識性能を比較する.

5.1 実験条件

Table 1 は, 収録した NAM データの特徴量抽出の分析条件を示す. 今回使用した HMM は 3 状態の left-to-right 型状態共有トライフォンモデルである. 状態数は 2189, 各状態の出力確率密度分布の混合数は 16 である. 認識時のデコーダは Julius4.01 [11] を用いており, 言語モデルは新聞記事から作成さ

Table 1 Analysis conditions

Sampling frequency	16 kHz
Window duration	25 ms
Frame shift period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12MFCCs, 12 Δ MFCCs, ΔE

れた CSRC2000 年度版の 6 万語の言語モデル [12] を使用する.

評価データとしては, 各話者が新聞記事を NAM で読み上げた 20 発話を用いており, 各話者の評価セットの難易度を揃えるために, パープレキシティ (PP) と未知語率 (OOVs) を可能な限り揃える. 全話者の評価セットの PP は, 平均で 48.23 であり, 最小値が 40.37, 最大値が 55.46 である. また, 全ての評価セットの OOVs は 0% である. 評価データ以外の NAM 発話 (1 話者 約 130~210 発話) を初期モデルの学習データ及び各話者の適応データとする. 適応時の初期モデルとして, 以下の 3 つのモデルを用いる.

1. 通常音声用不特定話者状態共有トライフォンモデル (**SI-Normal**)
2. 全話者の NAM データを用いて, **SI-Normal** の各パラメータを Baum-Welch アルゴリズムによって再学習したモデル (**SI-NAM**)
3. **SI-Normal** を元モデルとして, 全話者の NAM データを用いて SAT によって作成された正準モデル (**SAT-NAM**)

SI-Normal は, JNAS [13] の通常音声データを用いて学習したモデルである. **SAT-NAM** の学習において, CMLLR によって各話者への変換行列を推定する. この時, **SI-Normal** を作成する際の十分統計量を使用してクラス数 32 の回帰木を作成し, 1 話者につき複数の変換行列を推定する. また, CMLLR の推定を 1 回, 正準モデルの学習を 5 回反復するという処理を 3 回繰り返した. 各話者への MLLR 適応回数は 10 回とする.

5.2 実験結果

Table 2 は, 適応前後の全話者の平均単語正解精度を示している. **SI-NAM** における適応前の性能

Table 2 Word accuracy [%] averaged over all speakers before and after adaptation

Initial model	Before	After
SI-Normal	4.23	64.18
SI-NAM	53.25	68.61
SAT-NAM	26.44	73.17

は、NAM用の不特定話者音響モデルとしての性能を示している²。特定話者音響モデルの性能は、SI-NormalよりもSI-NAMを初期モデルとした時の方が優れている。そしてさらに、SAT-NAMを初期モデルとして用いることで性能が改善されている。最終的には、全話者の平均で73.17%の単語正解精度が得られている。また、適応後における全話者の単語正解精度の標準偏差を比較すると、SI-Normalを用いた時は14.81、SI-NAMを用いた時は12.63、SAT-NAMを用いた時は11.45である。これより、適応話者以外の話者のNAMデータも有効利用し、SATによって初期モデルを改良することで、平均的な認識率を向上させ、話者間の認識性能のばらつきを小さくする事が可能であることが分かる。

異なる初期モデルを用いることによって、適応後の認識性能がどのように変化するか更に詳細に考察する。SI-Normalを初期モデルとした場合と、SI-NAMあるいはSAT-NAMを初期モデルとした場合で、特定話者音響モデルの性能を比較し、結果をFig. 3に示す。Fig. 3は、SAT-NAMがSI-Normal、SI-NAMと比べてどの程度有効かを示している。SI-NAMは多くの話者でSI-Normalより認識率が上がっているが、悪影響を及ぼしている話者も多い。しかし、SAT-NAMを用いた場合は、ほぼ全ての話者で安定して認識率が向上している。このことから安定した初期モデルを構築する手法としてSATが有効であることが分かる。

またFig. 3より、全話者の単語正解精度の傾向は、どの初期モデルを使っても大きな差は見られない。つまり、SI-Normalを初期モデルとした際の特定話者音響モデルの性能が低い話者は、SAT-

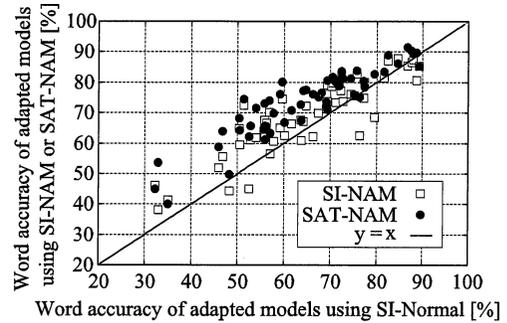


Fig. 3 Relationship of word accuracy of speaker adapted model for each speaker when using individual type of initial models

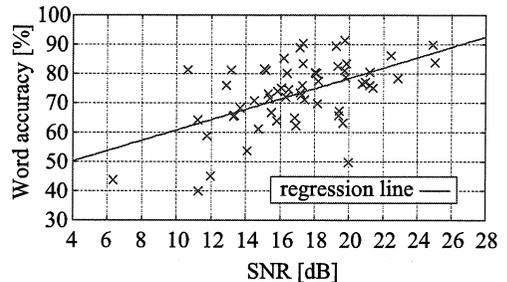


Fig. 4 Correlation between SNR and word accuracy of each speaker after adaptation using SAT-NAM. Correlation coefficient is 0.55.

NAMを用いることで、性能は向上するが、全話者の中では、性能が低いままである。

話者間の単語正解精度のばらつきの要因を調査するために、収録した各話者のNAMデータの平均SNRと認識率の相関を調査し、Fig. 4に示す。Fig. 4より、相関はそれほど高くなく、NAM認識の性能にはSNR以外にも、NAM発声の不安定さやNAM発声への慣れといった様々な要因が関わっていると考えられる。したがって、今後は更に詳細にNAMの音響特徴と認識率の関連性を調査する必要がある。

6 おわりに

本稿では、様々な話者のNAM認識において、事前収録したNAMデータを有効利用し、SATによって初期モデルを改良した際の特定話者音響モデルの性能の変化を報告した。適応後における全話者の平均単語正解精度がSATを用いることで大き

² ただし、学習データに評価話者のデータが含まれているので、話者に関してはクロスドテストである。

く改善された。また各話者の単語正解精度が従来の初期モデルに比べて、安定して改善されることが分かった。これらの結果より NAM 認識における SAT の有効性が実証された。また、話者毎の認識性能のばらつきを抑えることもできた。今後は、NAM の音響特徴と認識性能の関連性を詳細に調査し、さらに NAM 認識に適した音響モデル構築手法について検討する予定である。

謝辞 本研究の一部は、科研費基盤研究 A によって行われた。

参考文献

- [1] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, X. Huang, "Air-and-Bone Conductive Integrated Microphones for Robust Speech Detection and Enhancement", In Proc. ASRU, St. Thomas, U.S. Virgin Islands, pp. 249-254, 2003.
- [2] S.-C. Jou, T. Schultz, and A. Waibel, "Adaptation for Soft Whisper Recognition Using a Throat Microphone," In Proc. ICSLP2004, WeC2102p, 2004.
- [3] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session Independent Non-Audible Speech Recognition Using Surface Electromyography," In Proc. ASRU, pp. 331-336, 2005.
- [4] 中島淑貴, 鹿野清宏, "非可聴つぶやきをインタフェースとするコミュニケーションのためのソフトシリコーン型 NAM マイクロホンの開発," 電子情報通信学会論文誌 D, Vol. J89-D, No.8, pp. 1802-1810, August 2006.
- [5] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, K. Shikano, "Audible (Normal) Speech and Inaudible Murmur Recognition Using NAM Microphone," In Proc. EU-SIPCO2004, pp.329-332, 2004.
- [6] 長井孝之, 中村圭吾, 戸田智基, 猿渡洋, 鹿野清宏. "複数話者に対する非可聴つぶやき認識における各特定話者適応モデルの性能評価," 2008 年秋季音講論集, 1-1-7, pp.17-18, 2008.
- [7] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," In Proc. ICSLP96, vol.2, FrP2L1.3, 1996.
- [8] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs", Computer Speech and Language, Vol.9, pp. 171-186, 1995
- [9] C.H. Lee, C.H. Lin, and B.H. Juang, "A study on speaker adaptation of parameters of continuous density hidden Markov models," IEEE trans. Signal Processing., Vol. 39, pp. 806-814, 1991.
- [10] V.V. Digalakis, D. Rtishev, and L.G. Neumeier, "Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures," IEEE trans. Speech Audio Process., vol.3, no.5, pp.357-366, 1995.
- [11] A. Lee, T. Kawahara and K. Shikano. "Julius — an open source real-time large vocabulary recognition engine," In Proc. EU-ROSPEECH2001, pp. 1691-1694, 2001.
- [12] 河原 達也, 住吉 貴志, 李 晃伸, 武田 一哉, 三村 正人, 伊藤 彰則, 伊藤 克亘, 鹿野 清宏, "連続音声認識コンソーシアム 2000 年度版ソフトウェアの概要と評価.", 情報処理学会研究報告, 2001-SLP-38-6, pp.37-42, 2001.
- [13] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research," The Journal of the Acoustical Society of Japan, vol. 20, pp. 199-206, 1999.