音声の構造的表象を用いた未観測調音運動の推定に関する実験的検討*

☆内田秀継, 齋藤大輔, 峯松信明（東大）

1 はじめに

構音障がい者や語学学習者を対象にした発音トレーニングにおいて、調音運動情報を利用したトレーニング法が検討されている [1]。このトレーニング法では、学習者に対して、自らの現在の調音運動と目標とするべき調音運動の視覚的なフィードバックを与え、発音の誤りの対する直感的な理解を手助けする。調音フィードバックにおける学習者の現在の調音運動に関する情報は、調音観測システムによる測定や、音声－調音マッピングによる音声からの推定によって得ることができる。一方で、学習者が目標とするべき調音運動は、学習者がこれから獲得するものであるため、測定することは不可能である。さらに、音声－調音マッピングによる推定も、入力となる音声が存在しないため難しい。そこで、本稿では、当該話者による生成が困難な調音運動を，音声から推定するのではなく、音声の構造的表象から推定する方法について検討する。

2 音声の構造的表象を用いた未観測調音運動の推定

2.1 音声の構造的表象を用いた音声合成法

音声の構造的表象 [2] とは、音声を音響事象ごとに分布化し、その分布間の f-divergence（分布間距離）によって構造的に表したものである。分布間距離は、話者性（身体性）の違いや収録機器の違いといった静的な非言語的特徴に対して不変である。つまり、音声の構造的表象は、音声から話者性を削ぎ落とし、言語的特徴のみによって音声を記述したものである。

音声の構造的表象から音声を合成する手法 [3] では、言語的特徴のみを持つ音声の構造的表象に対して、発話者の身体性を付与することによって音声合成を実現する。音声の構造的表象に対する身体性の付与は、合成目標となる分布間距離を満たす音響事象を、発話者の音響空間内で探索することによって行われる。

合成目標となる音響事象（ターゲット事象）は、それと幾つかの音響事象（アンカー事象）との分布間距離によって表現されているとする。この時、アンカー事象群は、発話者の音響空間上の分布が事前に定義されているとする。このとき、ターゲット事象の音響空間内における探索は以下のコスト関数の最小化によって行われる。

$$J(C) = \frac{1}{2}\sum_{n=1}^{N}\left\{BD(C, C_n) - \eta_n\right\}^2 \qquad (1)$$

ここで、C,C_n は分布を表しており、特に C_n は n 番目のアンカー事象の分布である。また、η_n は他話者より得られるターゲット事象と n 番目のアンカー事象の分布間距離を表している。$BD(C_1, C_2)$ は分布 C_1,C_2 に関するバタチャリヤ距離である。式（1）は、分布 C とアンカー事象から計算される分布間距離と、ターゲット事象とアンカー事象の分布間距離の差を、分布 C に関するコスト関数として表したものである。このコスト関数の最小値を与える分布を、ターゲット事象の音響空間上の分布とする。

分布間距離は、話者非依存の特徴量であるため、任意の話者から発話から抽出したものを利用することができる。したがって、この手法では発話者がこれまで発話したことのない発音であっても、他者からその発音とアンカー事象となる発音の分布間距離を得ることで、発話者の音声でその発音を合成できる [4]。

2.2 未知観測調音運動の推定

音声の構造的表象からの音声合成法に基づき、未観測の調音運動運動を推定する方法について説明する。本稿では、推定対象となる未観測の調音運動に対して、その調音運動に対応する音声に関する分布間距離を手がかりとして推定を行う手法を提案する。

今、調音特徴量 $\boldsymbol{x}$ から音声特徴量 $\boldsymbol{y}$ への変換が、$\boldsymbol{y} = f(\boldsymbol{x})$ で表されるとする。この変換を用いて式（1）を以下のように改める。

$$J(\boldsymbol{x}) = \frac{1}{2}\sum_{n=1}^{N}\Big\{BD\Big(C(f(\boldsymbol{x}), \Sigma), C_n\Big) - \eta_n\Big\}^2 \qquad (2)$$

ここで、$C(f(\boldsymbol{x}), \Sigma)$ は、音響空間におけるガウス分布 $\mathcal{N}(f(\boldsymbol{x}), \Sigma)$ であり、その平均ベクトルは調音特徴量からの変換によって表されている。また、C_n は、音響空間におけるアンカー事象のガウス分布である。η_n は、他話者より得られるターゲット事象と n 番目のアンカー事象の音響特徴量の分布間距離を表している。式（2）は、推定対象となる調音運動に対して、その調音運動から得られる音声に関する分布間距離を拘束条件として与えたときのコスト関数である。そして、このコスト関数を最小化する調音特徴量を推定結果とする。この推定法では、発話者のアンカー事象における音響特徴量の分布 C_n、ターゲット事象とアンカー事象の音響特徴量に関する分布間距離 η_n、調音特徴量から音声特徴量への変換モデルが必要となる。

調音特徴量から音声特徴量への変換モデルとして、混合ガウス分布（GMM）による調音－音声マッピング [5] を用いる。音声特徴量を $\boldsymbol{y}$、調音特徴量を $\boldsymbol{x}$、二つの特徴量の結合ベクトルを $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$ とすると、結合ベクトルの確率分布は GMM によって以下のようにモデル化できる。

$$P(\boldsymbol{z}; \lambda^{(\boldsymbol{z})}) = \sum_{m=1}^{M}\alpha_m \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{\boldsymbol{m}}^{(z)}, \boldsymbol{\Sigma}_m^{(z)}) \qquad (3)$$

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(x,x)} & \boldsymbol{\Sigma}_m^{(x,y)} \\ \boldsymbol{\Sigma}_m^{(y,x)} & \boldsymbol{\Sigma}_m^{(y,y)} \end{bmatrix} \qquad (4)$$

ここで、m は混合成分のインデックス、α_m は混合成

---

*An experimental study of predicting unseen articulatory movements using speech structure, by Hidetsugu UCHIDA, Daisuke SAITO, Nobuaki MINEMATSU, (The University of Tokyo)

分の重みである。二乗誤差最小化基準（MMSE）による調音運動から音声へのマッピング関数は以下の式で表される。

$$\hat{\boldsymbol{y}} = \sum_{m=1}^{M} P(m|\boldsymbol{x};\lambda^{(\boldsymbol{z})})\{\boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}'_m(\boldsymbol{x} - \boldsymbol{\mu}_m^{(x)})\} \quad (5)$$

となる。ここで、$\Sigma' = \boldsymbol{\Sigma}_m^{(y,x)}\boldsymbol{\Sigma}_m^{(x,x)^{-1}}$ である。$p(m|\boldsymbol{x};\lambda^{(\boldsymbol{z})})$ を定数近似することで、このマッピング関数は、$\boldsymbol{x}$ に関する一次式となり、簡単な形式で式（2）に組み込むことができる。

3 実験

3.1 実験条件

音素を音響事象の単位として、単独の音素を対象とした調音運動の推定実験を行った。実験用データは、MOCHA-TIMIT [1] の男性話者の音声－調音パラレルデータを用いた。まず、コーパス内の全ての発話を用いて、各音素における音声特徴量と調音特徴量のガウス分布をそれぞれ求めた。このとき、分散共分散行列は対角行列とした。コーパス内において母音的特徴を持つ 20 個の音素に注目し、その中で任意の 1 音素を推定対象（ターゲット音素）、残りの 19 個の音素をアンカー音素として用いた。推定に用いるターゲット音素とアンカー音素の音声特徴量における分布間距離は、発話者の分布から計算した。実際の応用を考えると、分布間距離は発話者以外から求めるべきだが、分布間距離が話者非依存の特徴量をいうことを考えると、実験の妥当性を損なうものではない。調音－音声マッピングで用いる GMM は、コーパス内のデータからターゲット音素のデータを除いたもので学習した。音声特徴量は、MFCC（$C_1 \sim C_{24}$）とし、調音特徴量は、磁気センサシステムの受信センサの座標データ（14 次元データ）とした。ここで、調音特徴量に基づいて式（2）のコスト関数に以下で示す新たな拘束条件を追加する。

$$J'(\boldsymbol{x}) = J(\boldsymbol{x}) + \frac{1}{2}w_1\sum_{i=1}^{P}\sum_{k=1}^{D}(x_{i,k} - x_{i,k}^{(ave)})^2$$

$$+w_2\sum_{i=1}^{P}\sum_{k=1}^{D}\{\max(x_{i,k}^{(l)} - x_{i,k}, 0) + \max(x_{i,k} - x_{i,k}^{(u)}, 0)\}$$

$$+w_3\sum_{i=1}^{P}\sum_{j=i+1}^{P}\sum_{k=1}^{D}\max(|x_{i,k} - x_{j,k}| - d_{i,j,k}, 0) \quad (6)$$

ここで、$x_{i,k}$ は $\boldsymbol{x}$ の要素であり、i,k はそれぞれ測定点と座標の次元を表すインデックスである（$P = 7, D = 2$）。$x_{i,k}^{(ave)}$ は、アンカー音素の調音特徴量の各要素の平均値である。アンカー音素が母音の集まりであることから、第二項は、推定結果が平均的な母音の調音運動から逸脱することを制限するものとなっている。$x_{i,k}^{(l)}, x_{i,k}^{(u)}$ は、コーパス内のデータから求まる調音特徴量の各要素の下限と上限である。したがって、第三項は、推定結果が発話者の調音空間から逸脱することを制限するものとなっている。また、$d_{i,k,j}$ は、調音特徴量の各要素の任意のペア間の距離の上限を表しており、第四項は、磁気センサシステムの受信センサ同士の設置関係に関する制約となっている。$w_{1,2,3}$ は、それぞれの拘束条件に対する重みパラメータである。コスト関数 $J'(\boldsymbol{x})$ の $\boldsymbol{x}$ に関する最小化は、最急降下法を用いた。最急降下法の初期値は、ターゲット音素に対して最も小さい分布間距離を持つアンカー音素における調音特徴量の平均ベクトルとした。

[1] http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

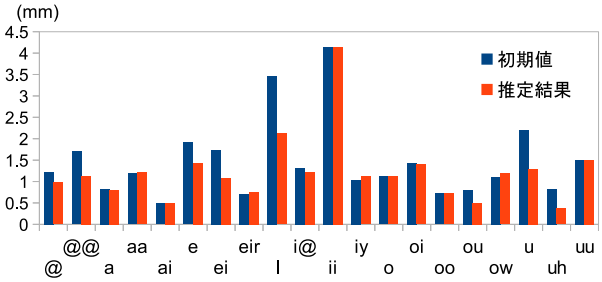

Fig. 1 Estimation errors of each target phoneme

3.2 結果

各ターゲット音素における推定誤差を Fig.1 に示す。推定誤差は、推定結果とターゲット音素の調音特徴量の平均ベクトルの各要素ごとの二乗誤差を平均したものである。提案法における推定結果と合わせて、最急降下法の初期値を推定結果とした場合の推定誤差を示している。音素の表記はコーパス内の表記に従った。初期値と推定結果を比べると、推定誤差がほとんど改善されていない音素もあるが、特定の音素においては推定誤差が大きく低下している。また、全音素について推定誤差を平均すると、初期値では 1.47mm、推定結果では 1.23mm であった。このことから、音声の構造的表象に基づいた推定は、未観測の調音運動の推定に有効であると言える。

4 おわりに

本稿では、未観測の調音運動の推定を、音声の構造的表象を利用して行う手法について検討した。実験の結果、提案法の有効性が示された。今回の実験では、推定対象は単独の母音音素としたが、今後は子音を含めた連続発声を対象とした実験を行っていく。

5 謝辞

本研究の一部は科研費・萌芽研究 (15K12059)、及び基盤研究 (A) (26240022) の助成を受けたものである。

参考文献

[1] A. Suemitsu, *et al.*, In *Proc. Acoustic society of Japan Autumn Meeting 2013*, pp.427-428, 2013.
[2] N. Minematsu, *et al.*, *Journal of New Generation Computing*, 28, 3, 299-319, 2010.
[3] D. Saito, *et al.*, *IEICE Technical Report*, SP2007-80, pp.55-60, 2007.
[4] R. Mihara, *et al.*, *IEICE Technical Report*, SP2009-71, pp.55-60, 2009.
[5] T. Toda, *et al.*, *Speech Commun*, vol. 50, pp.215-227, 2008.