Vol.2014-SLP-102 No.8
2014/7/25

MLPを用いた話者正準化に基づく音声認識の検討

久保田 雄一[1,a)]　大町 基[1]　小川 哲司[1]　小林 哲則[1]　新田 恒雄[1,2]

概要：不特定話者音声認識システムの性能向上を目的として，低演算かつ高精度な話者正準化手法を提案する．話者正準化の代表的な手法として，声道長正規化（Vocal Tract Length Normalization; VTLN）が広く用いられているが，計算量および表現能力の 2 つの観点で改善の余地がある．まず最適なワーピングパラメータを推定する際に，用意したパラメータ数だけ同時に音声認識を行う必要があるため，計算量が多くなる．また，VTLN は一つの発話内において同じパラメータで線形変換を実現している．しかし，最適な写像関数は音素ごとに異なると言われており，表現能力に改善の余地がある．そこで，本報告では多層パーセプトロン（Multi Layer Perceptoron; MLP）を用いた話者正準化手法を提案する．MLP は，任意話者の母音スペクトルを標準話者の母音スペクトルへ写像する関数を学習する．提案法は，(1) 認識時にパラメータを推定する必要がない (2) MLP により発話内で音素ごとに非線形な写像関数を実現させることができるという点で VTLN よりも優れる．しかしながら，スペクトルの低域および高域において歪が生じ，認識性能が低下する．この問題を解決するために，MLP による写像後のスペクトルと入力スペクトルの周波数重み付けを行う．不特定話者連続数字認識実験による評価では，提案法が VTLN と比較し 1.6 %性能を改善することを示す．

キーワード：音声認識，話者正準化，多層パーセプトロン，特徴抽出

A study on MLP-based speaker canonicalization

Abstract: Accurate and efficient speaker canonicalization is proposed to improve the performance of speaker-independent ASR systems. Vocal tract length normalization (VTLN) is often applied to speaker canonicalization in ASR; however, it requires parallel decoding of speech when estimating the optimal warping parameter. In addition, VTLN provides the same linear spectral transformation in an utterance, although optimal mapping functions differ among phonemes. In this study, we propose a novel speaker canonicalization using multilayer perceptron (MLP) that is trained with a data set of vowels to map an input spectrum to the output spectrum of a standard speaker or a canonical speaker. The proposed speaker canonicalization operates according to the integration of MLP-based mapping and identity mapping that depends on frequency bands and achieves accurate recognition without any tuning of mapping function during run-time. Results of experiments conducted with a continuous digit recognition task showed that the proposed method reduces the intra-class variability in both of the vowel and consonant parts and outperforms VTLN.

Keywords: Speech recognition, Speaker canonicalization, Multilayer perceptron, Feature extraction

1. はじめに

不特定話者音声認識では，話者間分散（音素毎の時間-周波数パターンの分散）が大きくなるため，性能が著しく低下する．性能改善のため，音響的なミスマッチを解決する話者適応 [1], [2]，入力スペクトルの時間-周波数パターンから話者不変な特徴量への変換を行う種々の話者正準化 [3], [4], [5] が提案されている．

標準的な話者正準化手法としては VTLN が広く用いられている．VTLN は，話者毎に異なる声道長の違いに対して異なる周波数ワーピング関数を用意する．一般的に，VTLN は複数のワーピングパラメータに対して並列で尤度計算を行うため，音声認識時に遅延を生じる．この問題に対し，江森らは，最尤法に基づき直接ワーピングパラメー

[1] 早稲田大学
Dept. of Compute Science, Waseda University, Tokyo, Japan
[2] 豊橋技術科学大学
Toyohashi University of Technology, Toyohashi, Japan
[a)] kubodo@pcl.cs.waseda.ac.jp

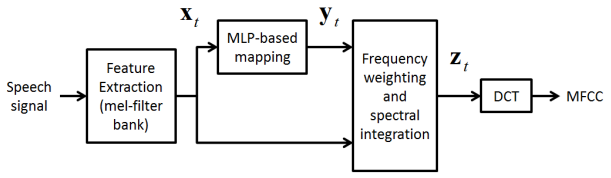

図 1 *Schematic diagram of proposed speaker canonicalization system*

タを推定することを試みている [6]．この方法は，発話を一単位とした線形写像により，話者間の声道長変動の影響を低減する．しかし，発話には異なる音素が複数含まれるため，声道の伝達関数は話者と音素双方の影響を受け，　最適なワーピングパラメータは発話内で変化する．そこで，ワーピングパラメータをフレームごとに選択する手法が提案されている [7], [8]．この手法では，ワーピングパラメータの選択により誤認識が低減するが，パラメータ探索に高い計算コストを必要とする．

本報告では，任意話者のスペクトルを標準話者のスペクトルへ変換する MLP に基づく話者正準化手法を提案する．提案法は，認識時の演算量・変換関数の表現能力の二つの点で VTLN に優れる．まず，提案法は，事前に MLP による写像関数パラメータを学習するため，認識時における写像関数の最適化が不要である．また，MLP は発話内で音素ごとに非線形な写像関数を学習できるため，VTLN で用いられる線形変換よりも高い表現能力をもつ写像関数を実現することができる．MLP に基づくスペクトル変換は，声質変換の分野で用いられている [9], [10]．これらの報告では，ケプストラム歪や聴覚的類似性に焦点を当てており，音声認識への適用と精度改善といった評価は行われていない．

一般に，話者の違いは母音部に現れると言われている．そこで本研究では，母音のデータセットを用いて MLP の写像関数を学習するアプローチを検討した．最初に行った予備実験では，MLP 写像後のスペクトルの低域と高域において歪が生じ，音声認識の性能を大きく低下させることが分かった．この問題に対処すべく，低域，中域，高域の帯域ごとに，MLP 写像の前後の二つスペクトルを重み付き加算することを試み，性能を大幅に改善できることを示す．

なお，本研究の成果は，話者変動に頑健な音声認識システムの開発にも活かすことができる．例えば，提案法を話者正規化学習 [11], [12] の枠組みに組み込むことも可能である．

以降 2. では，提案する話者正準化手法について説明するとともに，MLP に基づくスペクトル変換における周波数重みづけとスペクトル合成について述べる．次に 3. では連続数字音声認識実験の結果を報告し，4. で結論をまとめる．

2. MLP に基づく話者正準化

図 1 に提案する話者正準化手法の概要を示す．まず，スペクトル写像の対象となる標準話者を決定する．次に，任意話者のスペクトルパターンが標準話者のスペクトルパターンに変換されるように MLP を学習する．MLP の入出力特徴量は，対数メルフィルターバンクの出力を用いる．そして，図 1 に示すように，MLP 写像を行ったスペクトルと入力スペクトルの帯域別重み付けにより，正準化スペクトルを合成する．最後に，正準化スペクトルより MFCC を抽出する．以降では，2.1 で標準話者の決定法について述べ，2.2 で MLP に基づくスペクトル変換について述べる．そして 2.3 では MLP の入出力スペクトルに対する周波数重み付けを組み合わせた話者正準化について説明を行う．

2.1 標準話者の決定

話者が発声したスペクトルの対数メルフィルタバンク出力を，その話者を表す特徴量とする．複数話者の対数メルフィルタバング出力のうち，他話者の対数メルフィルタバンクの出力とのユークリッド距離が最小となる話者を標準話者として選択する．話者の特徴は主に母音に起因しているといわれている [13]．そこで，標準話者の選択には母音データのみを用いた．標準話者の決定には，JVPD コーパス [14] を用いた．このコーパスには，6–56 歳までの 385 話者（男性 186 人，女性 199 人）が発声した 5 母音が収録されている．まず，各話者の各母音を 1 つのセグメントとして切り出し，1 話者あたり 5 つのセグメントを抽出した．次に，各セグメントの中心フレームにおいて，24 次元の対数メルフィルタバンクの出力を抽出し，それを特徴ベクトル $\boldsymbol{x}_v^s$ とした．s と v はそれぞれ話者，母音を表す．最後に，以下に示す評価基準に基づき標準話者 S^* を決定した．

$$S^* = \arg\min_s \sum_{\forall s' \neq s} \sum_{\forall v} \left\| \boldsymbol{x}_v^s - \boldsymbol{x}_v^{s'} \right\| \tag{1}$$

その結果，身長 162.7 cm の 16 歳の女性が標準話者として選ばれた．

2.2 MLP に基づくスペクトル写像

話者 s が発声した音素 p の特徴ベクトル $\tilde{\boldsymbol{x}}^{(s,p)}$ と標準話者が発声した p の特徴ベクトル $\tilde{\boldsymbol{y}}^{(p)}$ の全音素に関する組み合わせ $\{(\tilde{\boldsymbol{x}}^{(s,p)}, \tilde{\boldsymbol{y}}^{(p)})\}_{\forall s}$ が既に与えられているものとすると，$\tilde{\boldsymbol{x}}^{(s,p)}$ から $\boldsymbol{y}^{(p)}$ への写像関数 $\tilde{\boldsymbol{y}}^{(p)} = f(\tilde{\boldsymbol{x}}^{(s,p)})$ を図 2 に示すような 1 つの MLP により表現することができる．VTLN では，音素に関わらず発話内共通の線形変換が行われている．一方，MLP に基づくスペクトル写像は，音素毎に異なる非線形写像を 1 つの MLP により行うことができる．そのため，VTLN より高い表現能力を実現することが可能と考える．

本報告では，$\tilde{\boldsymbol{x}}^{(s,p)}$ および $\tilde{\boldsymbol{y}}^{(p)}$ として，各音素の発声区間における中心フレームとその前後のフレームにおける 24 次元の対数メルフィルターバンクの出力である 3 フレームを連結したものを用いた．MLP の入力層，隠れ層，出力

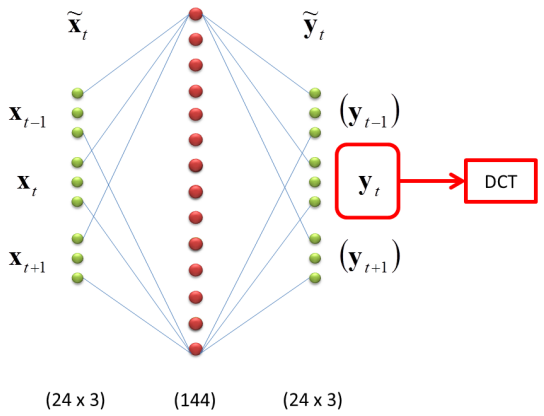

図 2 *Architecture of three-layered MLP for spectral mapping. In training, 24-dimensional logarithmic mel-filter bank channel outputs concatenated with previous and subsequent frames' outputs from arbitrary speakers and those from canonical speaker are applied to input and output layers. During run-time, 24-dimensional components ($\boldsymbol{y}_t$) in middle of 72-dimensional outputs ($\tilde{\boldsymbol{y}}_t$) are chosen as canonicalized filter bank outputs.*

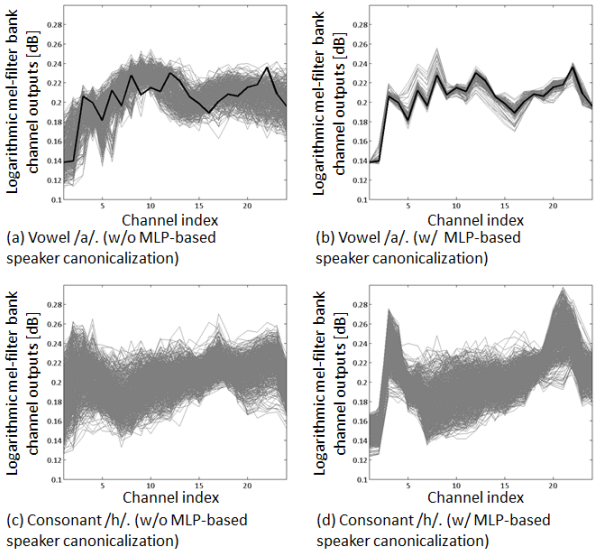

図 3 *Spectra with and without MLP-based spectral mapping. Bold line represents spectrum from canonical speaker. (a) and (c) show spectra of vowel /a/ and consonant /h/ without MLP-based mapping, respectively. (b) and (d) show spectra of /a/ and /h/ with MLP-based mapping, respectively.*

層におけるユニット数はそれぞれ 72，144，72 とした．また，2.1 で論じた理由により，MLP の学習に用いる音素は 5 母音のみとした．

認識時には，各フレームにおいて 24 次元の対数メルフィルタバンク出力 $\boldsymbol{x}_t \in \mathcal{R}^{24}$ を抽出し，その前後フレームを連結した 72 次元のベクトル $\tilde{\boldsymbol{x}}_t = (\boldsymbol{x}_{t-1}^{\mathrm{T}}, \boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{x}_{t+1}^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{R}^{72}$ を生成した．このベクトルを MLP の入力とし，MLP の出力 $\tilde{\boldsymbol{y}}_t \in \mathcal{R}^{72}$ を計算し，中心の 24 次元のベクトル $\boldsymbol{y}_t \in \mathcal{R}^{24}$ を抜き出した．$\tilde{\boldsymbol{y}}_t$ を離散コサイン変換することにより正準化した MFCC を求めた．

MLP の構造に関する予備実験として，入力層-隠れ層-出力層のユニット数が 72-144-24 または 72-144-72 の MLP を構築し，認識精度を調べた．その結果，72-144-72 の方が高い認識精度を示したため今後はそれを採用した．

2.3 周波数重みづけを用いたスペクトルの正準化

MLP に基づくスペクトル変換によって，ある周波数帯域でスペクトルに大きな歪が発生することがある．図 3 に MLP に基づくスペクトル変換適用前後の母音または子音のスペクトルを示す．これは leave-one-out で MLP 学習を母音毎に行ったものである．JVPD コーパスに含まれる 385 話者のうち，384 話者で MLP の学習，残りの 1 話者で評価を行った．図 3(a) および 3(b) より，任意話者から標準話者へのスペクトル変換により話者間の分散が減少している様子を確認することができる．しかし，標準話者（女性）が発声した母音の基本周波数が存在する 300 Hz 付近にスペクトルのピークがみられる．これは，スペクトルのピークが標準話者の基本周波数に向うように写像関数が学習されてしまったことが原因だと考えられる．また図 3(c) および 3(d) においては，スペクトル変換を行うことにより，子音のスペクトルの低域および高域においてが歪が生じる様子がみられる．この要因として，母音データのみを用いて MLP を学習していることが考えられる．

この問題を解決するために，MLP 写像後のスペクトルと入力スペクトルの周波数重み付けを行う．中域においては，スペクトル変換後のスペクトルを正準化スペクトルとして用い，低域および高域においては，スペクトル変換による悪影響が生じるため，入力スペクトルを正準化スペクトルとして用いる．この条件を満たすように，中域における信号のみを通過させるフィルタを MLP 写像後のスペクトルに適用し，低域と高域における信号のみを通過させるフィルタを入力スペクトルに適用する重みづけ関数を設計した．このフレームワークにより正準化された対数メルフィルターバンクの出力 $z_t(k)$ は，次式のように表すことができる．

$$z_t(k) = w^{\mathrm{out}}(k) \cdot y_t(k) + w^{\mathrm{in}}(k) \cdot x_t(k) \qquad (2)$$
$$w^{\mathrm{out}}(k) + w^{\mathrm{in}}(k) = 1.0 \quad \text{for} \quad \forall k$$

ここで，t および k はそれぞれフレーム番号 および 対数メルフィルタバンクのチャンネル番号を示し，$x_t(k)$ および $y_t(k)$ はそれぞれ MLP 写像前の対数メルフィルタバンク x_t および MLP 写像後の対数メルフィルタバンク y_t における k 番目の要素を示す．また，$w^{\mathrm{in}}(k)$，$w^{\mathrm{out}}(k)$ MLP 写像適用前後の対数メルフィルタバンクの k 番目の要素に対する重みを表す．図 4 に重みづけ関数の例を示す．重み

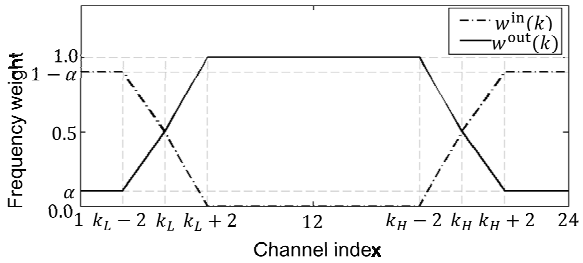

図 4 *Frequency weighting functions for inputs and outputs of MLP-based spectral mapping.*

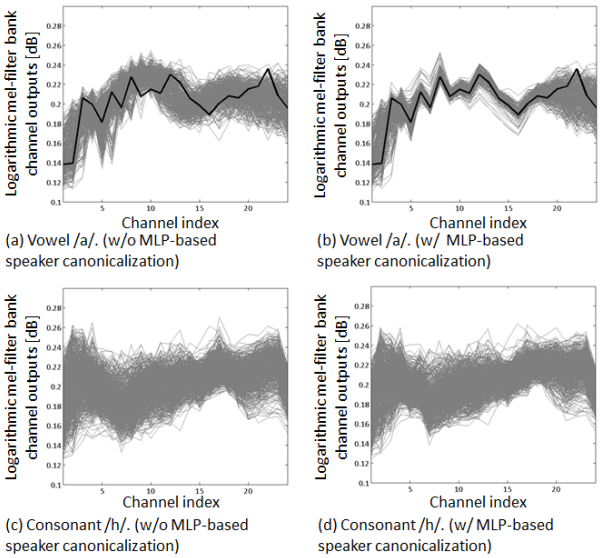

図 5 *Spectra with and without speaker canonicalizatoin applying frequency weighting and spectral integration to MLP-based mapping. Bold line represents spectrum from canonical speaker. (a) and (c) show spectra of vowel /a/ and consonant /h/ without spectral mapping, respectively. (b) and (d) show spectra of /a/ and /h/ with spectral mapping, respectively.*

付け関数は，4 つのパラメータ $\{\alpha,\ k_L,\ k_H,\ g\}$ により定義する．α は MLP 写像の高域および低域のチャンネルにおける出力の重み，k_L および k_H はそれぞれ上限周波数および下限周波数，g は k_L ， k_H における線形補間関数の傾きを示し，本研究では $g=(1-\alpha)/4$ とした．

図 5 に MLP 写像と周波数重み付けを組み合わせた変換を行う前後の母音および子音のスペクトルを示す．母音部の低域における基本周波数に起因するスペクトルピークや，子音部の低域および高域におけるスペクトルの歪を抑えられていることが確認できる．図 6 にスペクトル変換前後の母音部および子音部におけるスペクトルの分散を示す．MLP は母音のみで学習しているにも関わらず，MLP 写像により母音のスペクトルの分散だけでなく子音に関しても減らすことが出来た．一方，周波数重み付けにより，母音部における低域および高域の分散は増加してしまった．そのため，重み関数のパラメータは認識精度に応じて適切に設定する必要がある．

表 1 *Conditions for speech analysis*

sampling frequency	16 kHz
frame length	30 ms
frame shift	10 ms
type of window	Hanning window

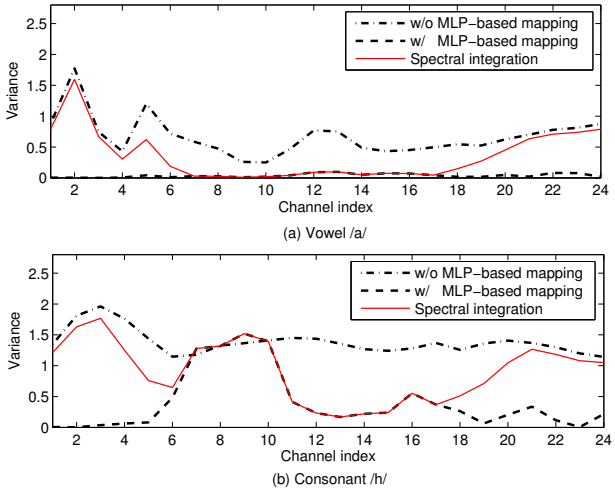

図 6 *Spectral variances of vowel /a/ and consonant /h/.*

3. 連続数字音声認識実験

連続数字音声認識タスクによる性能評価を行った．ここでは，以下の 3 つの特徴量を用いて数字正解精度を比較した．

(1) **MFCC** : 話者正準化未適用の MFCC
(2) **VTLN** : VTLN による話者正準化を適用した MFCC
(3) **MLP-FWSI** : 提案手法による話者正準化を適用した MFCC

3.1 実験条件

3.1.1 話者正準化

特徴抽出の条件を表 1 に示す．学習データとして，JVPD コーパスに含まれる 385 話者が発声した 5 母音のデータ，すなわち 1925 の母音データを用いた．385 話者のうち 1 人を正準化話者とした．MLP の各ユニットにおける活性化関数は傾き係数が 0.3 のシグモイド関数を用いた．学習はオンライン学習で行い，学習係数は 0.7 とした．図 4 に示したパラメータ k_L および k_H は，事前実験に基づき，それぞれ 5，19 とした．これは 400 Hz と 4000 Hz に相当する．また，$\alpha=0.0, 0.1, 0.2, 0.3, \cdots, 1.0$ とし，数字認識精度を評価した．

3.1.2 連続数字音声認識

認識実験で用いた音声データのサンプリング周波数および量子化ビットはそれぞれ 16 kHz および 16 bit で，フレー

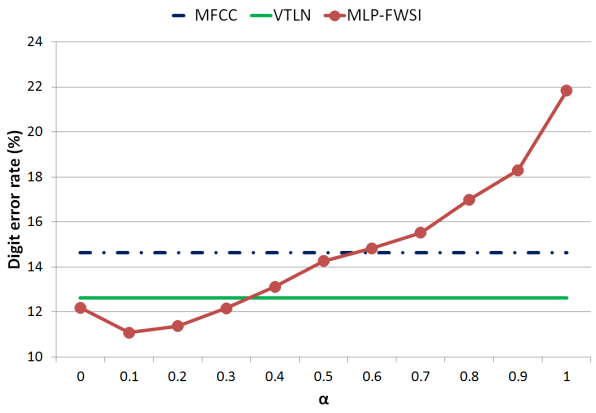

図 7 *Digit error rate (%) as a function of α, which represent weights for outputs from MLP-based mapping in low and high frequency channels.*

ム長 30 ms，フレームシフト 10 ms でフレーム分析を行った．音響特徴量は，12 次元の MFCC および Δ MFCC，パワーおよび Δ パワーを結合した 26 次元のベクトルを用いた．モデル学習および認識時には，発話ごとにケプストラム正規化を適用した．VTLN または提案手法は音響モデル学習評価の際に用いる全ての音声データすべてに適用した．

音響モデルには，性別非依存の monophone HMM を用いた．各状態の出力分布は 16 混合のガウス分布とし，対角共分散を用いた．学習には，日本語の新聞記事読み上げと音素バランス文で構成される JNAS [15] より選択した 41396 文を用いた．評価データには連続数字音声 (CENSREC-1) [16] より選択した 104 話者による 4004 発話を用いた．

3.2 実験結果

図 7 は連続数字認識システムから得られた数字認識誤り精度を表す．ここでは，提案手法において入力スペクトルに対する低域および高域における重みを制御するパラメータ α の違いによる認識精度の違いについても示した．$\alpha=0.0$ は MLP 写像の出力を高域と低域では用いないことを示し，$\alpha=1.0$ は全帯域において MLP 写像の出力を用いることを示す．$\alpha=1.0$ としたとき，話者正準化を適応していない **MFCC** に比べて数字認識誤り精度が低下した．一方，周波数毎に重み付けを行った **MLP-FWSI** は，$\alpha=0.1$ のときに最も高い精度を実現しており，数字認識誤り精度が，**MFCC** に比べて 3.6 %、**VTLN** に比べて 1.6 %減少した．この結果より，不特定話者の数字音声認識において，周波数重みづけを用いたスペクトル変換による提案手法を用いることにより，子音データを学習に用いることなく，高精度な話者正準化が行うことが可能であることがわかった．

4. まとめ

本報告では，MLP を用いた写像関数に基づく話者正準化法を提案した．この中で，MLP の出力単独ではスペクトルの高域と低域において大きな歪を生じることを示し，対処法として MLP の入出力スペクトルに対する周波数重み付けを組み合わせた話者正準化手法を提案し，この手法がクラス内分散を大きく低減できることを示した．また，連続数字音声認識タスクに対する評価では，標準的な **MFCC/HMM** と話者正準化を行った **VTLN–MFCC/HMM** に対して，提案手法が数字認識精度で各々 3.6 %と 1.6 %向上することを示した．今後は提案の話者正準化手法の改善とともに，LVCSR のタスクへ適応した際の評価を行いたい．

謝辞 本研究は SCOPE の補助による．

参考文献

[1] G. Saon, H. Soltau, D. Nahamoo and M. Picheny, “Speaker adaptation of neural network acoustic models using I-Vectors,” IEEE Works. on ASRU, pp. 55–59, Dec 2010.
[2] O. Abdel-Hamid, H. Jiang, “Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition.” Proc. Interspeech2013, pp.1248-1252, 2013.
[3] E. Eide and H. Gish, “A parametric approach to vocal tract length normalization,” Proc. ICASSP, pp.346-348, May 1996.
[4] H. Wakita, “Normalization of vowels by vocal-tract length and its application to vowel identification,” IEEE Trans. ASSP, vol.25, no.2, pp.183–192, April 1997.
[5] L. Welling, S. Kanthak, and H. Ney, “Improved method for vocal tract normalization,” Proc. ICASSP, pp.346-348, May 1996.
[6] T. Emori and K. Shinoda, “Rapid vocal tract length normalization using maximum likelihood estimation,” Proc. Eurospeech2001, pp.1649-1652, Sept. 2001.
[7] A. Miguel, E. Lleida, R. Rose, L. Buera, and A. Ortega, “Augmented state space acoustic decoding for modeling local variability in speech,” Proc. Interspeech2005, pp.3009-30, Sept. 2005.
[8] F. Müller and A. Mertins, “Enhancing vocal tract length normalization with elastic registration for automatic speech recognition,” Proc. Interspeech2012, Sept. 2012.
[9] S. Desai, A. W. Black, B. Yenanarayana and K. Prahallad, “Spectral mapping using artificial neural networks for voice conversion,” IEEE Trans. on ASLP, vol. 18, no. 5, pp. 954–964, July 2010.
[10] N. W. Ariwardhani, Y. Iribe, K. Katsurada and T. Nitta, “Voice conversion for arbitrary speakers using articulatory movement to vocal-tract parameter mapping,” Proc. MLSP, pp. 1–6, 2013.
[11] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, “A compact model for speaker-adaptive training,” Proc. ICSLP, vol.2, pp.1137–1140, Oct. 1996.
[12] M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” Computer Speech & Lang., vol.12, no.2, pp.75–98, April 1998.
[13] M. J. Owren and G. C. Cardillo, “The relative roles of vowels and consonants in descriminating talker identity versus word meaning,” J. Acoust. Soc. Am., vol.119, no.3, pp.1727–1739, 2006.
[14] Vowel database: Five Japanese vowels of males, females, and children along with relevant physical data (JVPD):

http://research.nii.ac.jp/src/en/JVPD.html.
[15] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Kobayashi, K. Shikano, and S. Itabashi, “The design of the newspaper-based Japanese large vocabulary continuous speech recogonition corpus,” Proc. ICSLP, pp.3261-3264, Nov. 1998.
[16] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, and T. Endo, “AURORA-2J: An evaluation framework for Japanese noisy speech recognition,” IEICE Trans. Inf. & Syst., vol. E88-D, no. 3, pp.535-544, March 2005.