# Dynamic frequency warping に基づくスペクトル変換を用いた非ネイティブ音声の発音変換*

◎北条伸克, 亀岡弘和，田中宏，金子卓弘 (NTT)

## 1 はじめに

本稿では，音声の話者性を保ちながら，訛りの情報を変換する，発音変換に取り組む．訛りとは，国や地域，社会集団などの共同体による，発音の違いである [1]．訛りは，フォルマントやその軌跡 [2, 3]，イントネーションや話速のパラメータ [4, 1] の違いとして現れることが知られている．話者性を保ちながら，訛りのない音声へ変換することが可能となれば，第二言語学習者のための言語教育システム [5, 6, 7] や，様々な国籍の参加者が利用する会議システムなどのアプリケーションにとって有用である．

声質変換とは，言語情報を保ちながら話者性などに代表される非言語情報を変換する技術である．訛りを一つの非言語情報と見なせば，発音変換は声質変換の特別な問題とみなすことができる．声質変換では，混合ガウスモデル（Gaussian mixture models; GMMs) に基づく手法 [8, 9] など，統計的手法が成功を収めている．このような声質変換の手法を発音変換に適用した場合，入力話者の話者性が目標話者のものへ変換される点が問題となる．これは，訛りの変換と話者性の変換が，一つの音響特徴量のマッピング関数により同時に表現されるためである．

本研究は，dynamic frequency warping (DFW) に基づくスペクトル変換がこの問題を解決できるという仮説に基づく．DFW に基づくスペクトル変換は，主に声質変換の自然性を向上させる目的で過去に提案されている [10, 11]．音声の訛りはフォルマント周波数軌跡の違いに現れるため [2, 3], 周波数ワーピングにより，周波数軸上で各フォルマントをネイティブ話者のフォルマントの位置に配置されれば，発音の変換が可能であると期待される．また，DFW はフォルマントパワーやスペクトル傾斜を変換しないため，入力話者の話者性は大きな影響を受けないと期待される．

本研究の目的は，DFW に基づくスペクトル変換が，発音変換に対し有用であるかを検討することである．本研究では，さらに 2 点について検討を行う．一つは，差分スペクトルモデルの有用性の検討である．DFW は，入力スペクトルを周波数方向のみに変換するため，スペクトルのパワー方向の差を減少させることができない．フォルマントのパワーも発音の品質にとって重要な要因であると考えられるため，スペクトルパワーの変換により，発音を目標話者のものに近づけることが可能であると考えられる．しかし，入力スペクトルを目標話者のものに変換すると，入力話者の話者性が保持できない．したがって，発音品質の変換と話者性の保持にはトレードオフがあると考えられる．本研究では，このトレードオフについて，主観評価実験による評価を行う．二つ目は，周波数ワーピング関数のモデル化方法である．先行研究 [11] では，まず入力話者と目標話者のスペクトルの結合分布を GMM によりモデル化し，各ガウシアンに周波数変換関数を対応づける．この周波数変換関数は，各ガウス関数に対応するフレームにおける入力話者，目標話者のスペクトルの平均スペクトルから得られる．しかし，平均スペクトルは過剰に平滑化されるため，得られる周波数変換関数もまた過剰に平滑化する傾向がある．この問題を回避するため，本研究では，フレームごとに入力話者と目標話者のスペクトルから周波数変換関数を抽出し，得られた周波数変換関数を予測する特徴量とみなし，入力スペクトルと周波数変換関数のパラメータの結合分布を GMM によりモデル化する．本研究では，発音変換の問題のうち，スペクトル変換のみを扱う．

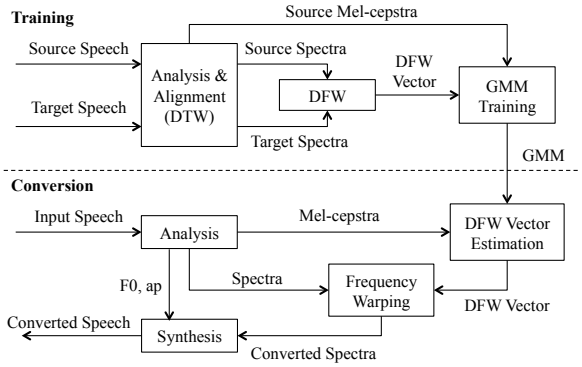



Fig. 1: 提案法の構成．

## 2 Dynamic Frequency Warping に基づく発音変換

提案法は，学習部と変換部から構成される．提案法の構成を図 1 に示す．

### 2.1 周波数差分を用いた Dynamic frequency warping

提案法は，まずパラレルデータの各フレームで，DFW [12, 13] により，最適な周波数変換関数を推定する．時間同期の取られた入力スペクトルを $\boldsymbol{X} = \{\boldsymbol{x}_t\}_{t=1}^{T}$，目標スペクトルを $\boldsymbol{Y} = \{\boldsymbol{y}_t\}_{t=1}^{T}$ で表す．ここで，$t$, $f$ はそれぞれ，時刻，周波数のインデックスを表す．次式により得られる，変換スペクトルと目標スペクトルの距離を最小化する変換関数 $\hat{\boldsymbol{w}}_t = [\hat{w}_{f,t}]_{f=1}^{F}$ を，本稿では DFW ベクトルと呼ぶ．

$$\hat{\boldsymbol{w}}_t = \underset{w_1,\cdots,w_F}{\arg\min} \sum_{f=1}^{F} \mathcal{D}(x_{w_f,t}, y_{f,t}) \qquad (1)$$

*Automatic speech pronunciation correction of non-native speech with dynamic frequency warping-based spectral conversion by Nobukatsu HOJO, Hirokazu KAMEOKA, Kou TANAKA and Takuhiro KANEKO (NTT)

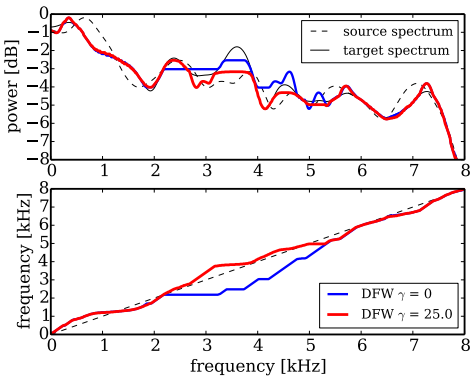



Fig. 2: DFW により変換されたスペクトル（上）と抽出された DFW ベクトル（下）の例．式（2）において，$\gamma = 0$，$\gamma = 25.0$ とした結果をそれぞれ青，赤で示した．

ここで，$w_f$ は周波数インデックスの値 $w_f \in \{1, \cdots, F\}$ を昇順に取るものとする．推定されるワーピングの経路を，各 $f \in \{1, \cdots, F-1\}$ で $w_{f+1} - w_f \in \{0, 1, 2\}$ となるよう制約することで，滑らかな，固定長の DFW ベクトルを抽出することが可能である．$\mathcal{D}(x, y)$ として，対数スペクトルの $l_2$ ノルムを用いることも可能であるが，得られる変換スペクトルはプラトーを持つ傾向にある．プラトーは，入力スペクトルと目標スペクトルのパワーの差に起因する．プラトーは，高調波を劣化させ，変換音声の品質を劣化させる懸念があるため，本研究では，スペクトル距離関数に，周波数差分の項を導入する．

$$\mathcal{D}(x_f, y_f) = ||\log x_f - \log y_f||_2 + \gamma ||\dot{x}_f - \dot{y}_f||_2, \quad (2)$$

$$\dot{x}_f = \log x_{f+1} - \log x_f, \quad (3)$$

$$\dot{y}_f = \log y_{f+1} - \log y_f \quad (4)$$

ここで，$\gamma$ は周波数差分項の重みを表す．周波数差分項の導入は，相関に基づく DFW [14] やスペクトルピークのヒストグラムに基づく DFW [15] と同様の効果があると期待される．周波数差分を用いた周波数変換の例を図 2 に示した．周波数差分項の導入により，プラトーの少ないスペクトルが得られることがわかる．

また，話者性の情報が DFW ベクトルに含まれるのを避けるため，DFW ベクトルの抽出を行う前に，声道長正規化（vocal tract length normalization; VTLN）[16] を行うことが有用であると考えられる．

### 2.2 DFW に基づくスペクトルの変換

提案法では，入力スペクトルと DFW ベクトルの結合分布を GMM によりモデル化する．特徴量ベクトルの次元を圧縮するため，結合ベクトルを $\boldsymbol{z}_t = [\boldsymbol{m}_t^{\mathrm{T}}, \boldsymbol{d}_t^{\mathrm{T}}]^{\mathrm{T}}$ のように構成する．ただし，$\boldsymbol{m}_t$ は入力スペクトル $\boldsymbol{x}_t$ から抽出されるメルケプストラムである．ベクトル $\boldsymbol{d}_t$ は次式により得られる．

$$\boldsymbol{d}_t = \mathrm{DCT}(\boldsymbol{w}_t - \boldsymbol{w}_b) \quad (5)$$

ただし，$\boldsymbol{w}_b = [1, \cdots, F]^{\mathrm{T}}$ であり，DCT$(\cdot)$ は離散コサイン変換を表す．結合ベクトルは，次式のように，GMM によりモデル化される．

$$p(\boldsymbol{z}) = \sum_{i=1}^{I} \alpha_i N(\boldsymbol{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \sum_{i=1}^{I} \alpha_i = 1, \alpha_i > 0, \quad (6)$$

ただし，$N(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均ベクトル $\boldsymbol{\mu}$，共分散行列 $\boldsymbol{\Sigma}$ により表現される正規分布である．$\alpha_i$ はクラス $i$ の重み，$I$ はガウス関数の混合数を表す．$\{\boldsymbol{m}_t, \boldsymbol{d}_t\}_{t=1}^{T}$ を訓練データとして与えられたとき，EM アルゴリズムにより GMM を学習する．

変換時には，学習された GMM と入力スペクトルを用いて，DFW ベクトルが推定される．マッピング関数は，次式により与えられる [9]．

$$F(\boldsymbol{m}) = E[\boldsymbol{d}|\boldsymbol{m}] = \sum_{i=1}^{I} h_i(\boldsymbol{m})[\boldsymbol{\mu}_i^{(d)} + \boldsymbol{\Sigma}_i^{(dm)}(\boldsymbol{\Sigma}_i^{(mm)})^{-1}(\boldsymbol{m} - \boldsymbol{\mu}_i^{(m)})] \quad (7)$$

$$h_i(\boldsymbol{m}) = \frac{\alpha_i N(\boldsymbol{m}; \boldsymbol{\mu}_i^{(m)}, \boldsymbol{\Sigma}_i^{(mm)})}{\sum_j^I \alpha_j N(\boldsymbol{m}; \boldsymbol{\mu}_j^{(m)}, \boldsymbol{\Sigma}_j^{(mm)})}, \quad (8)$$

ここで，$\boldsymbol{\mu}_i^{(m)}$ と $\boldsymbol{\mu}_i^{(d)}$ は，$\boldsymbol{\mu}_i$ のうち，$\boldsymbol{m}$ と $\boldsymbol{d}$ に対応する次元のベクトルである．DFW ベクトルは，次式により推定される．

$$\tilde{\boldsymbol{w}}_t = \mathrm{iDCT}(F(\boldsymbol{m}_t)) + \boldsymbol{w}_b, \quad (9)$$

ただし，iDCT$(\cdot)$ は逆離散コサイン変換を表す．推定された $\tilde{\boldsymbol{w}}_t$ は連続値を取るため，四捨五入により整数に変換した上で，変換スペクトル $\tilde{\boldsymbol{y}}_t = [x_{\tilde{w}_{f,t},t}]_{f=1}^{F}$ を得る．

### 2.3 差分スペクトルモデル

DFW は，入力スペクトルを周波数方向のみに変換するため，スペクトルのパワー方向の差を減少させることができない．フォルマントのパワーも発音の品質にとって重要な量であると考えられるため，スペクトルパワーの変換により，発音を目標話者のものに近づけることが可能であると考えられる．しかし，入力スペクトルを目標話者のものに変換すると，入力話者の話者性が保持できない．したがって，発音品質の変換と話者性の保持にはトレードオフがあると考えられる．本研究では，このトレードオフについて，主観評価実験による評価を行う．本節では，変換スペクトルと目標話者のスペクトルのパワーの差分（差分スペクトル）を予測し，変換スペクトルに差分スペクトルを加算する手法を提案する．

差分スペクトル $\boldsymbol{r}_t = [r_{f,t}]_{f=1}^{F}$ は，目標話者のスペクトルと変換スペクトルの差分として定義される．

$$r_{f,t} = \frac{y_{f,t}}{x_{\tilde{w}_{f,t},t}} \quad (10)$$

差分スペクトルを用いて，結合ベクトル $\boldsymbol{s}_t = [\boldsymbol{m}_t^T, \boldsymbol{q}_t^T]^{\mathrm{T}}$ を構成する．ここで，$\boldsymbol{q}_t$ は $\log \boldsymbol{r}_t$ の DCT を表す．結合ベクトルは，別途 GMM によりモデル化され，DCT ベクトルの推定と同様に，差分スペクトル $\tilde{\boldsymbol{r}}$ の推定に使用する．変換スペクトル $\tilde{\boldsymbol{y}}_t^{(r)} = [\tilde{y}_{f,t}^{(r)}]_{f=1}^{F}$ は次式により得られる．

$$\tilde{y}_{f,t}^{(r)} = \tilde{y}_{f,t} \cdot \tilde{r}_{f,t}^{\lambda} \quad (11)$$

ここで，$\lambda$ は差分スペクトルモデルの重みを表す．

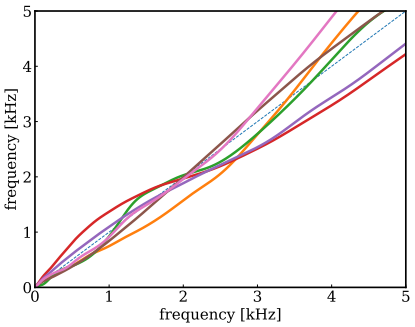



Fig. 3: $\boldsymbol{\mu}_i^{(d)}$ から再構成した DFW ベクトルの例．簡単のため，16 のガウス関数のうち 6 つについてプロットした．

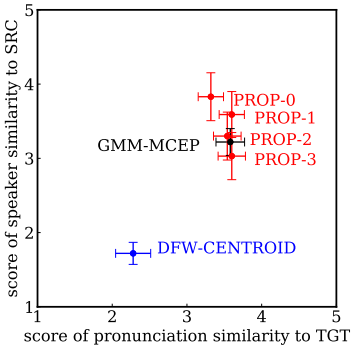



Fig. 4: 主観評価実験結果．

## 3 実験

### 3.1 実験条件

従来法，提案法による変換音声の，目標話者に対する発音の類似性と，入力話者に対する話者性の類似性を評価した．本実験では，インド人の男性入力話者（以下，SRC）と，アメリカ人の男性目標話者（以下，TGT）の音声データを使用した．音声データは，65 発話（約 5 分）を使用した．音声データのうち 20 発話は評価データとして使用し，残りを訓練データとして使用した．サンプリング周波数は 16kHz であり，STRAIGHT 分析 [17] により，5ms ごとに 25 次元のメルケプストラムと基本周波数（$F_0$），非周期性指標を抽出した．入力話者と目標話者の特徴量系列は，DTW [18] による時間同期を取った．本実験では，下記の 6 手法について評価を行った．

- GMM-MCEP: メルケプストラム特徴量と GMM による従来の声質変換手法. [9].
- DFW-CENTROID: DFW に基づくスペクトル変換手法.
- PROP-0: 差分スペクトルモデルを使用しない提案手法 ($\lambda = 0.00$).
- PROP-1: 差分スペクトルモデルを使用する提案手法 ($\lambda = 0.33$).
- PROP-2: 差分スペクトルモデルを使用する提案手法 ($\lambda = 0.67$).
- PROP-3: 差分スペクトルモデルを使用する提案手法 ($\lambda = 1.00$).

DFW-CENTROID は DFW に基づく声質変換手法 [11] を模倣して実装したが，音声波形生成のためのボコーダ等，詳細は異なる．提案法について，発音品質の向上と話者性の劣化のトレードオフを評価するため，差分スペクトルモデルの重みを 4 種類に設定し，評価を行った．

DFW-CENTROID は，まず，入力話者と目標話者のスペクトルから線スペクトル対（line spectral pairs; LSP) 特徴量を抽出し，その結合ベクトルを GMM によりモデル化した．続いて，各ガウシアンについて，ガウス関数の平均ベクトルから，入力話者と目標話者の平均スペクトルの対を再構成した．入力話者と目標話者の平均ベクトルの対から，2.1 節の DFW 手法を適用することで周波数変換関数を得た．スペクトル変換方法は，メルケプストラムではなく LSP を使用する点を除いて，提案法と同一のアルゴリズムを使用した．LSP の次元は 25 とした．

提案法では，25 次元のメルケプストラムを用いた．また，DFW ベクトルは DCT により，25 次元に圧縮した．低周波数領域で高い周波数分解能を得るため，DFW ベクトルを抽出する際は，メルスペクトルを使用した．

本実験では，VTLN を実施せずに DFW ベクトル抽出を行った結果を掲載した．VTLN の実施により実験結果が改善しなかったためである．本実験で使用した音声データでは，SRC と TGT の声道長が同程度であったためであると考えられる．

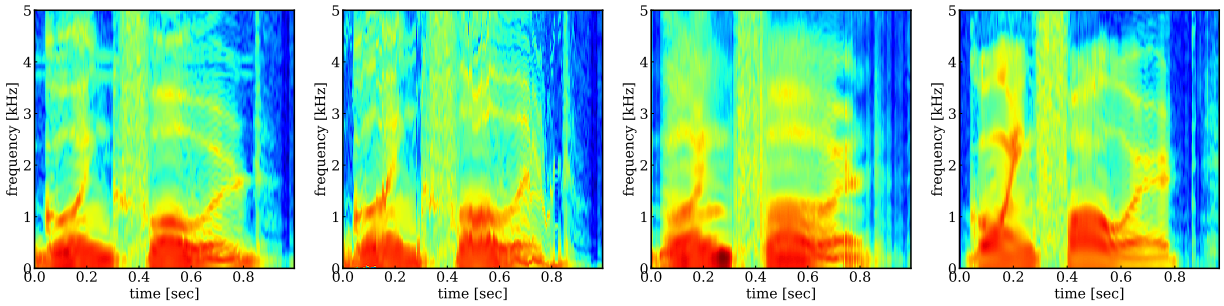
全ての実験条件で，GMM の混合数は 16 に設定した．各手法で変換されたスペクトルと，SRC の音声から抽出された F0 および非周期性指標を使用して，STRAIGHT [17] により変換音声の波形を生成した．

PROP-0 について，学習された GMM の $\boldsymbol{\mu}_i^{(d)}$ から再構成された周波数変換関数の例を図. 3 に示した．$1-3$ kHz の周波数領域において，200 Hz 程度の変換幅を持つ関数が学習されていることが確認できる．

### 3.2 主観評価実験

TGT に対する発音の類似性と，SRC に対する話者の類似性を 5 段階の DMOS（5: 非常に似ている ~1: 非常に似ていない）により評価した．被験者は 10 名とした．

主観評価実験結果を，5%信頼区間とともに図. 4 に示した．まず，GMM-MCEP と PROP-0 の結果を比較すると，GMM-MCEP が高い発音類似性を示すことが明らかとなった．この結果から，差分スペクトルモデルを使用しない提案法は，従来の声質変換手法に比べ，発音の類似性で劣ることが明らかとなった．しかし，GMM-MCEP と PROP-1 を比較すると，発音類似性は同程度である一方で，話者類似性は PROP-1 が上回ることが明らかとなった．この結果から，差分スペクトルモデルを使用する提案法は，従来の GMM に基づく声質変換と同程度に発音を変換する効果がある一方で，入力話者の話者性を保つことが可能であることが明らかとなった．次に，DFW-CENTROID と PROP-0 を比較すると，発音類似性，話者類似性とも，PROP-0 が上回ることが明らかとなった．この結果から，DFW ベクトルを特徴量系列として扱い，統計モデル化する提案法が有用であることが示された．また，従来法間の結果を比較すると，PROP-1 の発音類似性は PROP-0 を上回るものの，PROP-2，PROP-3




(a) SRC (b) PROP-1 (c) GMM-MCEP (d) TGT

Fig. 5: SRC, PROP-1, GMM-MCEP, TGT による，発話 “Going forward.” のスペクトログラムの例．フォルマント周波数の変化を明示するため，0 - 5 kHz の周波数帯域を示した．

は PROP-1 と同程度の発音類似性であり，話者類似性は劣化することが明らかとなった．以上から，今回の実験条件では，差分スペクトルモデルに対する 0.67 以上の重みは有効出ないことが明らかとなった．

評価データ中の文章 “Going forward.” に対する SRC, PROP-1, GMM-MCEP，TGT のスペクトログラムの例を図. 5 に示した．

## 4 まとめ

本稿では，話者性を保ちながら訛りのない音声を変換する，発音変換のため，DFW に基づくスペクトル変換の有用性を検証した．主観評価実験により，DFW に基づく発音変換と差分スペクトルモデルを併用することで，従来の GMM に基づく声質変換に比べ，同程度の発音類似性と，より高い話者類似性を実現可能であることを示した．今後の課題として，他の話者，言語での評価を行う．

## 参考文献

[1] Q. Yan and S. Vaseghi, “A comparative analysis of uk and us english accents in recognition and synthesis,” in *Proc. ICASSP 2002*, vol. 1. IEEE, 2002, pp. I–413.

[2] J. Harrington, F. Cox, and Z. Evans, “An acoustic phonetic study of broad, general, and cultivated australian english vowels,” *Australian Journal of Linguistics*, vol. 17, no. 2, pp. 155–184, 1997.

[3] C. I. Watson, J. Harrington, and Z. Evans, “An acoustic comparison between new zealand and australian english vowels,” *Australian journal of linguistics*, vol. 18, no. 2, pp. 185–207, 1998.

[4] J. C. Wells, *Accents of English*. Cambridge University Press, 1982, vol. 1.

[5] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, “Foreign accent conversion in computer assisted pronunciation training,” *Speech Communication*, vol. 51, no. 10, pp. 920 – 932, 2009, spoken Language Technology for Education. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639308001763

[6] D. Felps, C. Geng, and R. Gutierrez-Osuna, “Foreign accent conversion through concatenative synthesis in the articulatory domain,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2301–2312, 2012.

[7] D. Felps and R. Gutierrez-Osuna, “Developing objective measures of foreign-accent conversion,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1030–1040, 2010.

[8] Y. Stylianou, “Applying the harmonic plus noise model in concatenative speech synthesis,” *IEEE Transactions on speech and audio processing*, vol. 9, no. 1, pp. 21–29, 2001.

[9] 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.

[10] T. Toda, H. Saruwatari, and K. Shikano, “Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum,” in *Proc. ICASSP 2001*, vol. 2. IEEE, 2001, pp. 841–844.

[11] D. Erro, A. Moreno, and A. Bonafonte, “Voice conversion based on weighted frequency warping,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[12] H. Valbret, E. Moulines, and J.-P. Tubach, “Voice transformation using psola technique,” *Speech communication*, vol. 11, no. 2-3, pp. 175–187, 1992.

[13] N. Maeda, B. Hideki, S. Kajita, K. Takeda, and F. Itakura, “Speaker conversion through non-linear frequency warping of straight spectrum,” in *Sixth European Conference on Speech Communication and Technology*, 1999.

[14] X. Tian, Z. Wu, S. W. Lee, and E. S. Chng, “Correlation-based frequency warping for voice conversion,” in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. IEEE, 2014, pp. 211–215.

[15] E. Godoy, O. Rosec, and T. Chonavel, “Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[16] J. Cohen, T. Kamm, and A. G. Andreou, “Vocal tract normalization in speech recognition: Compensating for systematic speaker variability,” *The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3246–3247, 1995.

[17] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight,” in *Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2001.

[18] Y. Stylianou, “Statistical methods for voice quality transformation,” *Proc. European Speech Communication Association, Madrid, Spain, 1995*, pp. 447–450, 1995.