Vol.2017-MUS-115 No.23
2017/6/17

フルバンド音声を対象とした音声分析合成システムに用いるスペクトル包絡の音質劣化のない低次元表現

宮下 玄太[1,a)]　森勢　将雅[1]　小澤 賢司[1]

概要：本稿ではフルバンド音声分析合成系を対象としたスペクトル包絡の音声符号化について述べる．本研究におけるフルバンド音声とは，可聴周波数範囲を満たすナイキスト周波数を持つ，サンプリング周波数 40 kHz 以上の音声である．音声符号化に関する研究は，一般にはサンプリング周波数 16 kHz 以下の低域音声に焦点を当てて行われてきた．一方で，現在の統計的パラメトリック音声合成では，フルバンド音声のパラメータの低次元表現が使用されている．本研究は，フルバンド音声を対象とし，スペクトル包絡のメルケプストラム解析に焦点を当てて行った．メルケプストラム解析で利用される周波数軸の伸縮には 3 種類 (mel, Bark $\mathrm{ERB_N}$) の聴覚スケールを使用した．音声分析合成系 WORLD を用いて得られたスペクトル包絡を対象に符号化を行い，作成した合成音声を用いた主観評価実験により，最適な次元数は約 40 であることを示した．聴覚スケールの種類は，合成音の音質に有意な影響を与えないことも確認した．

キーワード：音声分析合成，音声符号化，スペクトル包絡，周波数スケーリング変換

1. はじめに

統計的音声合成 (SPSS: Statistical parametric speech synthesis)[1] は広く研究されている分野であり，近年では主に Deep neural network (DNN)[2] を用いた合成が行われている．DNN を用いた合成は，ボコーダを用いてパラメータ化された，基本周波数 (F0)，スペクトル包絡，非周期性指標の 3 つの音声パラメータを利用してモデルの構築を行う [3]．DNN モデルを構築するためのツールキットである Merlin[4] は，STRAIGHT[5] と WORLD[6] 等を利用している．生のオーディオ波形を生成するための DNN である WaveNet[7] は音声の波形を直接モデル化することが出来るが，トレーニングの段階で F0 が必要となる．自動音声認識 (ASR) では，波形を用いた音響モデルが提案されている [8]．ASR と高品質の音声合成の違いは音声のサンプリング周波数にあり，ASR は主に 16 kHz 以下のサンプリング周波数の狭帯域音声を使用しているが，高品質音声合成はフルバンド音声を主に使用する．フルバンド音声を使用して直接波形を処理することは現段階では困難である．

SPSS は，音声のスペクトル情報を効率よく学習するため，Deep Auto–Encoder を用いたスペクトル構造の低次元特徴抽出 [9] が提案されている．これは，高速フーリエ変換 (FFT) で計算されたパワースペクトルを入力として，高品質のボコーダによって推定されたスペクトル包絡と比較して高品質の音声を合成することが可能である．

Auto–encoder による音響特徴抽出の性能はトレーニングに使われるデータに依存するので，本研究ではメルケプストラム解析に基づく信号処理によって，学習データに依存しないスペクトル包絡の低次元表現に焦点を当てる．本研究の目的は，フルバンド音声について

- 符号化せずに音声合成された音声を自然に合成するための適切な次元数
- メルケプストラム解析における周波数スケーリング変換が音質に及ぼす影響

の 2 点を示すことである．これらの情報は，適切な次元の数が SPSS としてだけでなく，低次元表現を抽出するためのベースラインとしても使用が可能であるため，DNN を用いた音声合成にも利用することが可能となる．

本論文の第 2 節では，音声符号化に関する関連研究について論じ，主観評価の概要を述べる．第 3 節では実験条件を表した後，実験結果を示す．第 4 節では，その結果について考察を行い，適切な次元数と周波数スケーリング変換のモデルを示す．第 5 節では，全体の要約と今後の研究について述べる．

[1] 山梨大学
[a)] g17tk022@yamanashi.ac.jp

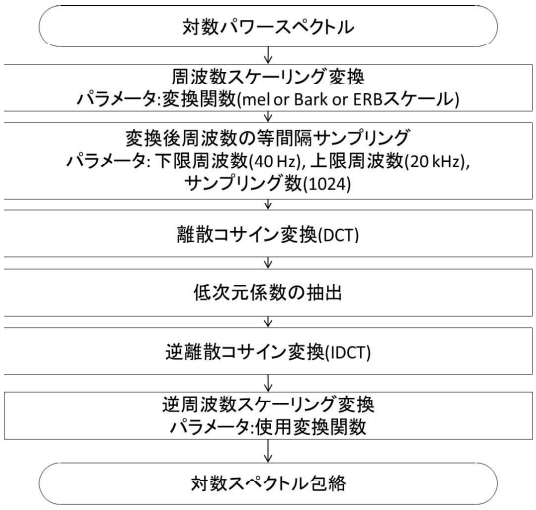

図 1　スペクトル包絡を符号化する処理の流れ

2. 音声符号化過程

2.1 符号化技術

狭帯域音声の符号化は古くから研究が行われている．線形予測符号 (LPC: Linear predictive coding)[10] は，最も使われているアルゴリズムの 1 つであり，線スペクトル対 (LSP: Line spectral paris)[11] は通信システムにおいて広く用いられる．ケプストラム [12] はいくつかの改良されたアルゴリズムの基盤となっている．まず，一般化ケプストラム [13] が提案され，その後にメルケプストラム解析 [14], [15] が行われた．

メル一般化ケプストラム [16] は，音声合成の研究で広く使用されている．メル一般化ケプストラムは，周波数スケーリング変換のためのパラメータを持ち，狭帯域音声のスペクトルの非線形伸縮を行う．フルバンド音声を用いる SPSS では，Warped Linear Prediction (WLP) の研究等が行われている [17].

メルケプストラム解析において，MLSA フィルタ (mel–log spectrum approximation filter)[18] はメルケプストラムから音声波形を直接合成する．一方で，ボコーダを用いた合成では，パラメータ化されたスペクトル包絡から音声を合成することが可能である．本実験ではボコーダを用いて，スペクトル包絡に対し聴覚スケールに基づいた周波数スケーリング変換を行う．符号化の目的としては，フルバンド音声分析合成のための劣化のないスペクトル包絡を得ることとする．

2.2 メルケプストラム解析

図 1 に提案するスペクトル符号化の概要とそのパラメータを示す．周波数スケーリング変換では，聴覚スケールに基づく 3 種類の変換法がある．

mel スケール [19] は，ピッチの知覚尺度のうち最も一般的なものの 1 つである．このスケールは 2 つの係数を持ち，複数の係数の組み合わせが提案されているが，本研究では式 (1) を採用した．これは，最も一般的な係数 [20] から 1 つを利用した．f は入力周波数を示す．

$$\mathrm{mel}(f) = 1127.01048 \log\left(\frac{f}{700} + 1\right) \tag{1}$$

Bark スケール [21] とは，ラウドネスの主観的測定に関連する音響心理学的スケールであり，式 (2) に基づいて算出される．逆関数を求めることが困難であるので，式 (2) から近似解を用いて線形に補間することで逆関数を求める必要がある．

$$\mathrm{Bark}(f) = 13 \arctan(0.00076f) + 3.5 \arctan\left(\left(\frac{f}{7500}\right)^2\right) \tag{2}$$

$\mathrm{ERB_N}$(Equivalent Rectangular Bandwidth) スケール [22] も，音響心理学的スケールであり，人間の聴覚で行われる音声の周波数スペクトル解析を再現したモデルである．式 (3) によって定義される．

$$\mathrm{ERB_N}(f) = 21.4 \log\left(\frac{4.37f}{1000} + 1\right) \tag{3}$$

3 つの変換関数は，対数変換と近いが，主な違いは低周波数帯域にある．図 2 にその違いを示す．縦軸は 24 kHz の周波数で 1.0 を示すように正規化したものである．この 3 つの変換関数は，それぞれのスケーリングで周波数の分解能を決定する．

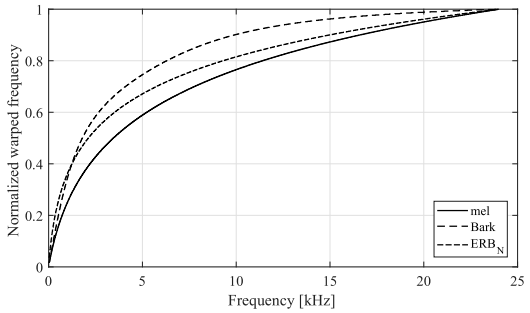

図 2　聴覚スケールに基づく 3 種類の変換関数比較．縦軸は 24 kHz で 1.0 となるよう正規化している．

2.3 周波数スケーリング変換

変換後のスペクトルは，周波数スケーリング変換によって非線形に伸縮された周波数軸上から等間隔でサンプリングされる．この変換後スペクトルには下限周波数，上限周波数，サンプル数という 3 つのパラメータが関係する．本研究では，下限周波数を一般的な F0 推定の下限周波数に従って 40 Hz と定める．また，上限周波数を人間の可聴周波数範囲の上限である 20 kHz と定める．サンプル数は周波数伸縮を行った後，復号可能なスペクトル包絡の最大次

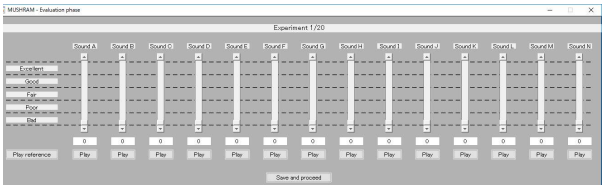

図 3　MUSHRA 法による主観評価実験に用いた GUI.

元数となる．本研究では，高品質音声分析合成システムである WORLD (D4C edition[23]) を使用しており，これはフルバンド音声に対して 2048 の FFT ポイント数を用いている．標本化定理より，サンプル数は 2048 の半数である 1024 とした．伸縮された周波数軸におけるサンプルは離散周波数 bin とは一致しないため，線形補間を行うことにより周波数の値を求めた．

2.4　スペクトル包絡の低次元表現

符号化の手順としては，まずサンプリングされた対数スペクトル包絡に対して，離散コサイン変換 (DCT) を行い，次に低次元抽出を行う．抽出する次元数は設定可能なパラメータであり，それに基づいて符号化効率が決定される．次元を N に設定した場合，スペクトル包絡は $N/1025$ に圧縮される．抽出された係数は，伸縮された周波数軸上で逆 DCT によって対数スペクトル包絡に変換される．伸縮されたスペクトル包絡は，使用された変換関数の逆関数を用いて線形の周波数軸上に再変換される．

3.　主観評価

符号化を行っていないスペクトル包絡と比較して劣化のないスペクトル包絡の低次元表現をする際に，適切な次元数を決定するため主観評価実験を行う．周波数スケーリング変換関数の差異も同様に評価する．

3.1　実験用ボコーダ

本実験では，高品質のボコーダとして，WORLD (D4C edition[23]) を用いて音声分析合成を行った [4]．各音声パラメータ推定には，DIO[28], CheapTrick[26], [27], D4C[23] を使用し，それぞれ F0，スペクトル包絡，非周期性指標を求めた．音声パラメータからの合成には MLSA フィルタでなく，WORLD の合成関数を使用した．F0 推定における周波数の下限及び上限は，デフォルト設定として 71 及び 800 Hz と設定した．フレーム幅は 5 ms に設定した．F0 の推定結果を視覚的に確認し，有声区間が無声区間と誤って識別された箇所を修正した．また，再合成された音声の音質において，実験を行う際に致命的な欠陥がないことを確認した．スペクトル包絡に使用される FFT ポイント数は 2048 と設定した．その他のパラメータはデフォルト設定とした．

表 1　実験条件

評価手法	
比較方法	MUSHRA 法
被験者	20 代男性 12 名
実験環境	防音室 (A-weighted SPL: 18 dB)
オーディオインターフェース	Roland QUAD-CAPTURE
ヘッドフォン	SENNHEISER HD650
使用音声	親密度別単語了解度試験用音声データセット 2007(FW07)
評価用音声	
発話者	4 人 (男女各 2 名)
A/D 変換	48 kHz/16 bit
音源数	全 20 音声 (各発話者 5 音声)
音声種	4 モーラ単語
周波数変換条件	
聴覚スケール	mel，Bark，$\mathrm{ERB_N}$
次元数	20，30，40，50
下限/上限周波数	40/20000 Hz
サンプリング数	1024

3.2　実験条件

表 1 に主観評価の条件を示す．実験の評価方法として MUSHRA 法 (Method for the subjective assessment of intermediate quality levels of coding systems) を用いた．MUSHRA 法とは，提示された音声の品質を評価する方法の 1 つで，被験者は図 3 に示す GUI を用いて 0 から 100 の尺度で音声刺激を採点する．この手法は，一般的に音質の違いの評価に使用されていて，MOS 評価よりも差の検出力が高い手法と言われている．Non significant (n.s.) は，音声が同じ音質を持っていることを保証するものではなく，MUSHRA 法の検出力においてその差が検出されなかったことを意味する．

実験は，暗騒音の A-weighter SPL が 18 dB の防音室を使用し，通常の聴力を有する 12 人が評価に参加した．音声刺激はヘッドホン (SENNHEISER HD650) を用いて与えた．主観評価に用いた音声刺激は，2 人の男性と 2 人の女性による各 20 発話である．サンプリング周波数は 48 kHz，量子化ビット 16 bit である．発話内容は日本語による子音を含む 4 モーラ単語であり，親密度別単語了解度試験用音声データセット 2007 (FW07) を使用した [29]．

被験者は，同時に提示される 14 種類の音声刺激を順に評価する．14 個の内訳は，符号化されたスペクトル包絡を用いて合成された 12 個の音声刺激と，2 個の音声刺激 (元音声と符号化なし WORLD 合成音) となっている．12 個の音声刺激は，3 種類の周波数スケーリング変換関数 (mel, Bark，$\mathrm{ERB_N}(f)$) と，4 種類の次元 (20, 30, 40, 50) の積となる．これらの次元数は，予備実験にて決定した．

3.3　実験結果

図 4 に実験結果を示す．縦軸は各条件の MUSHRA 法による評点，誤差棒は 95 %信頼区間を示す．まず，実験結果について統計分析を行う．考察のために複数の比較が必

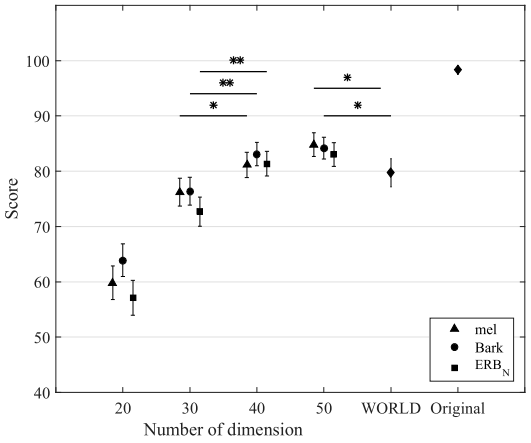

図 4　MUSHRA 主観評価結果．*は有意差 $p < 0.05$ を示す．**は有意差 $p < 0.01$ を示す．

要であるため，Benjamini–Hochberg 法 [30] に基づく 2 段階線形上昇手順 [31] を実施した．

表 2 は，複数の比較リストを示す．カッコ内の値は各周波数スケーリング変換関数で使用される次元数を表す．補正後の p 値が基準値を超える場合，補正後の p 値の代わりに n.s. を示す．この結果より，30 と 40 次元の差は全ての聴覚スケールにおいて有意差があり，40 と 50 次元の差に有意差はなかった．同じ次元における聴覚スケールの比較では，30，40 および 50 において有意差は検出されなかった．50 次元の低次元表現と WORLD を比較した際，mel スケールと Bark スケールには有意差が検出された．この有意差は低次元表現が WORLD より音質が優れていることを示している．その理由は次節にて考察を行う．

本実験の結果をまとめると，以下の 2 点になる．

- 次元数は約 40 次元にて符号化前と劣化なく音質が維持できる．
- 30 以上の次元において聴覚スケールによる差に有意差はなかった．

次節では，低次元表現が WORLD よりも音質評価値が高かった理由について分析し，考察を行う．

4. 考察

まず，低次元表現が符号化なしのスペクトル包絡より優れた音質を持つ理由を説明する．その後，従来手法と比較した際の符号化効率について論ずる．

4.1 符号化による音質向上現象

実験結果より，50 次元の mel スケールと Bark スケールは WORLD の音質評価値を明確に上回っていた．図 5 は実験結果を各話者別にまとめたものである (50 次元 mel スケールと WORLD のみ)．図 5 より，女性 A 及び男性 A にのみ，音質評価値が WORLD よりも低次元音の方が高くなる現象の有意差が認められた．以下では，この原因について解析した結果について述べる．

表 2　検証する組み合わせとその補正 p 値．括弧内の値は次元数を表す．補正 p 値が基準値を超える場合は n. s. (non significant) と表す．

組み合わせ	補正 p 値 [32]
WORLD，mel(50)	0.020
WORLD，Bark(50)	0.030
WORLD，ERB(50)	n.s.
mel(40)，mel(50)	n.s.
Bark(40)，Bark(50)	n.s.
ERB(40)，ERB(50)	n.s.
mel(30)，mel(40)	0.0206
Bark(30)，Bark(40)	0.0008
ERB(30)，ERB(40)	0.00002
mel(30)，Bark(30)	n.s.
mel(30)，ERB(30)	n.s.
Bark(30)，ERB(30)	n.s.
mel(40)，Bark(40)	n.s.
mel(40)，ERB(40)	n.s.
Bark(40)，ERB(40)	n.s.
mel(50)，Bark(50)	n.s.
mel(50)，ERB(50)	n.s.
Bark(50)，ERB(50)	n.s.

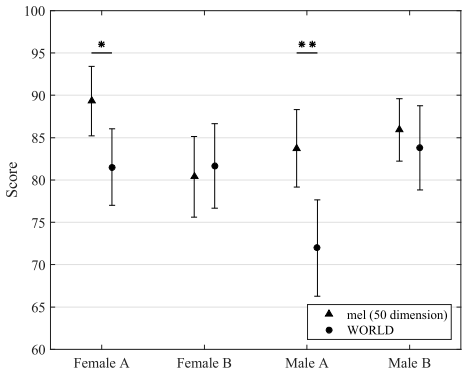

図 5　WORLD と mel50 次元評点の発話者別比較．2 人の話者について有意差がある．

4.2 フレーム間差分の分析

WORLD の評点が 50 次元音よりも低くなる原因の調査の一環として，フレーム間差分に関する音声の定量分析を行った．周波数帯域別の影響を確認するため，全体の周波数 24 kHz を 4.8 kHz 間隔で 5 分割したスペクトログラムの時間差分 $\Delta S(n,k)$ の分析を行った．スペクトログラムの時間差分 $\Delta S(n,k)$ は，フレーム間のパワースペクトルの差分を示したものである．これは式 (4) によって定義される．

$$\Delta S(n,k) = \sqrt{(10\log S(n+1,k) - 10\log S(n,k))^2}. \quad (4)$$

この数式において，n はフレーム番号，k は離散周波数，$S(n,k)$ はスペクトログラムを表す．$\Delta S(n,k)$ を全フレームについて求めることで，各音声の時間ごとの平均的なパワースペクトルの変動量を表すことが可能である．変動量

$\mathrm{var}(k)$ は式 (5) によって定義される．

$$\mathrm{var}(k) = \frac{\sum_{n=0}^{N} -1\Delta S(n,k)}{N-1}. \quad (5)$$

この数式において，N は音声のフレーム数を表す．各フレームの離散周波数は 1025 次元で表されているため，この次元数の分割を行った．分割は 4.8 kHz 毎に行っている．分割後の変動量 $\mathrm{var_d}(l)$ は式 (6) によって定義される．

$$\mathrm{var_d}(l) = \sum_{k=205l+1}^{205(l+1)} \frac{\mathrm{var}(k)}{205} \ (l = 0, 1, \cdots, 4). \quad (6)$$

この数式において，l は分割番号を表す．次元数の分割は，スペクトログラムにおける変化を確認した際に，最も符号化時に差が現れる帯域に限定して平滑化量を調査する目的で行った．平滑化量とは，フレーム間のパワースペクトルの差が，WORLD に比べて低次元表現がどの程度減少しているかを示す．平滑化量 $\mathrm{diff}(l)$ は，式 (7) によって定義される．

$$\mathrm{diff}(l) = \mathrm{rav}(l)_{\mathrm{world}} - \mathrm{rav}(l)_{\mathrm{com}} \quad (l = 0, 1, \cdots, 4). \quad (7)$$

この数式において，$\mathrm{var}(l, N)_{\mathrm{world}}$ は WORLD の変動量を示し，$\mathrm{var}(l, N)_{\mathrm{com}}$ は 50 次元表現の変動量を示す．ここでは，全発話音声を対象に平滑化量 $\mathrm{diff}(l)$ を計算し，平滑化量 $\mathrm{diff}(l)$ と，WORLD と 50 次元表現の評点差分の相関分析を行った．結果を図 6 に示す．結果として，4.8 kHz から 9.6 kHz の帯域においての平滑化量 diff(2) と，WORLD と 50 次元表現の評点差分との間に，相関係数 0.699 の相関が見られた．WORDL と 50 次元表現の評点差分は，全被験者の平均評価を，各発話音声別に算出した．この相関の有意差検定を行ったところ，p <0.01 であった．これらより，フレーム間におけるスペクトル包絡の大きな変動が品質低下の原因となっており，それが符号化によりスペクトル包絡が平滑化され，影響が軽減されたと考えられる．また，平滑化量は低次元表現の評点が，WORLD よりも高いほど多くなる傾向にあった．

4.3 符号化効率と従来法比較

Deep Auto–encoder による低次元表現 [9] では，フルバンド音声について，59 次元にスペクトル包絡を圧縮することが可能である．この研究では，F0 と非周期性指標はそれぞれ 1 次元と 25 次元で表されているため，次元数は 1 フレームあたり 85 次元となる．文献 [23] によれば，非周期性指標は 5 次元に圧縮が可能であることが示されている．本研究結果より，スペクトル包絡の次元数を 40 次元に設定した場合，1 フレームあたり 46 次元 (1 次元 F0，40 次元スペクトル包絡，5 次元非周期性指標) に圧縮することが可能である．本研究による低次元表現をベースラインとして Deep auto–encoder を使用することによって，さらなる圧縮も期待される．

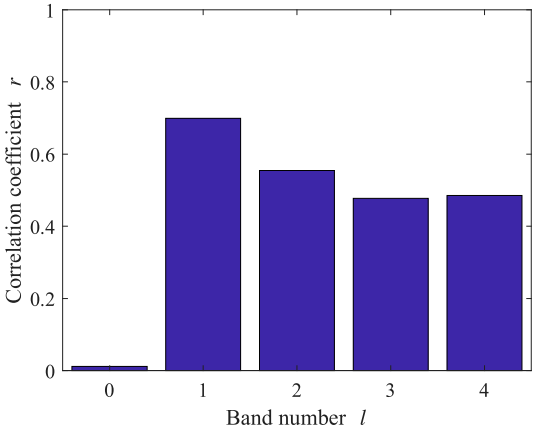

図 6 分割番号による相関係数比較．

5. 結論

本研究では，メルケプストラム解析を用いて，聴覚スケールに基づいた周波数スケーリング変換を行った．フルバンド音声を用いた主観評価を行い，スペクトル包絡を表す際に必要な次元数と変換関数を決定した．その結果，40 次元程度で，符号化なしの WORLD による合成音声と同程度の音質を持つことを示した．また，40 次元において変換関数の種類は音質に影響を与えなかった．

今後の発展としては，音声パラメータにおける量子化と，適切なフレーム幅の設定が挙げられる．F0 が 200 Hz のとき，基本周期は 5 ms であるため，F0 が 200 未満の音声では 5 ms 分析で十分であると考えられる．しかし，女性発話において F0 は 200 Hz を超過することがある．これは，声帯振動が 5 ms より短い周期で生じることを意味する．よって，5 ms 以下の周期での分析により音質が改善する可能性がある．これらの定量的に最適化されたパラメータは，音声分析合成システムだけでなく，フルバンド音声を使用する SPSS 分野で利用することが出来る．

6. 謝辞

本研究の一部は，科研費 JP15H02726, JP16H05899, JP16K12511, 16H01734 の支援を受けて行われた．

参考文献

[1] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” Speech Communication, vol. 51, pp. 1039 –1064, 2009.
[2] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in Proc. ICASSP2013, pp. 7962–7966, 2013.
[3] H. Dudley, “Remaking speech,” J. Acoust. Soc. Am., vol. 11, no. 2, pp. 169–177, 1939.
[4] Z. Wu, O. Watts, and S. King, “Merlin: An open source neural network speech synthesis system,” in Proc. of SSW 2016, pp. 218–223, 2016.

[5] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, “Re–structuring speech representations using a pitch–adaptive timefrequency smoothing and an instantaneou–frequency–based F0 extraction,” Speech Communication, vol. 27, no. 3–4, pp. 187–207, 1999.
[6] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder–based high–quality speech synthesis system for real–time applications,” IEICE Trans. Inf. Syst., vol. E99–D, pp. 1877–1884, 2016.
[7] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499, 2016.
[8] Z. Tuske, P. Golik, R. Schluter, and H. Ney, “Acoustic modeling with dpee neural networks using raw time signal for LVCSR,” in Proc. INTERSPEECH 2014, pp. 890–894, 2014.
[9] S. Takaki and J. Yamagishi, “A deep auto–encoder based low–dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis,” in Proc. ICASSP2016, pp. 5535–5539, 2016.
[10] B. Atal and S. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” J. Acoust. Soc. Am., vol. 50, no. 2B, pp. 296–302, 1971.
[11] F. Itakura, “Line spectrum representation of linear predictor coefficients of speech signals,” J. Acoust. Soc. Am., vol. 57, no. S1, p. S35, 1975.
[12] A. Oppenheim and R. Schafer, “Homomorphic analysis of speech,” IEEE Trans. Audio and Electroacoust., vol. AU–16, no. 2, pp. 221–226, 1968.
[13] K. Tokuda, T. Kobayashi, and S. Imai, “Generalized cepstral analysis of speech –unified approach to LPC and cepstral method,” in Proc. ICSLP-90, pp. 37–40, 1990.
[14] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel–cepstral analysis of speech,” in Proc. ICASSP92, vol. 1, pp. 137–140, 1992.
[15] K. Tokuda, “Speech coding based on adaptive melcepstral analysis,” in Proc. ICASSP’ 94, pp. 197–200, 1994.
[16] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, “Mel–generalized cepstral analyasis – a unified approach to speech spectral estimation,” in Proc. ICSLP–94, pp. 1043–1046, 1994.
[17] T. Raitio, A. Suni, M. Vainio, and P. Alku, “Wideband parametric speech synthesis using warped linear prediction,” in Proc. INTERSPEECH2012, pp. 1420–1423, 2012.
[18] S. Imai, “Mel log spectrum approximation (MLSA) filter for speech synthesis,” Electronics and Communications in Japan (Part I Communications), vol. 66, no. 2, pp. 10–18, 1983.
[19] S. Stevens, J. Volkmann, and E. Newman, “A scale for the measurement of the psychological magnitude pitch,” J. Acoust. Soc. Am., vol. 8, no. 3, pp. 185–190, 1937.
[20] S. Stevens and J. Volkmann, “The relation of pitch to frequency: A revised scale,” The American Journal of Psychology, vol. 53, no. 3, pp. 329–353, 1940.
[21] E. Zwicker and H. Fastl, “Analytical expressions for critical-band rate and critical bandwidth as a function of frequency,” J. Acoust. Soc. Am., vol. 68, no. 5, pp. 1523–1525, 1980.
[22] B. Moore, Psychology of Hearing. Academic Press, 2003.
[23] M. Morise, “D4C, a band–aperiodicity estimator for high–quality speech synthesis,” Speech Communication, vol. 84, pp. 57–65, 2016.
[24] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, “TANDEM–STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation,” in Proc. ICASSP2008, pp. 3933–3936, 2008.
[25] H. Kawahara and M. Morise, “Technical foundations of TANDEM–STRAIGHT, a speech analysis, modification and synthesis framework,” SADHANA – Academy Proceedings in Engineering Sciences, vol. 36, no. 5, pp. 713–728, 2011.
[26] M. Morise, “CheapTrick, a spectral envelope estimator for high–quality speech synthesis,” Speech Communication, vol. 67, pp. 1–7, 2015.
[27] –, “Error evaluation of an f0-adaptive spectral envelope esti- mator in robustness against the additive noise and f0 error,” IEICE Trans. Inf. Syst., vol. E98–D, no. 7, pp. 1405–1408, 2015.
[28] M. Morise, H. Kawahara, and H. Katayose, “Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech,” in Proc. AES 35th International Conference, CD–ROM, pp. CD–ROM, 2009.
[29] Kondo, T., Amano, S., Sakamoto, S., and Suzuki, Y.:Development of familiarity–controled word–lists (fw07). IEICE Society Conference research report, Vol. 107, No. 432, pp. 43–48 (2008).
[30] Y. Benjamini, A. M. Krieger, and D. Yekutieli, “Adaptive linear step-up procedures that control the false discovery rate,” Biometrika, vol. 93, no. 3, pp. 491–507, 2006.
[31] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” J. R. Statist. Soc. B, vol. 57, no. 1, pp. 289–300, 1995.
[32] S. P. Wright, “Adjusted p-values for simultaneous interface,” Biometrics, vol. 48, pp. 1005–1013, 1992.