Vol.2016-SLP-112 No.2
2016/7/28

巨大特定話者データを用いた HMM・DNN・RNNに基づく音声合成システムの性能評価

Xin Wang[1,a)]　高木 信二[2,b)]　山岸 順一[2,1,c)]

概要：本論文では男性 100 時間，女性 50 時間の巨大特定話者データを用い統計的パラメトリック音声合成システムの構築を行い，学習データ量の違いによる性能への影響を調査する．近年，統計的パラメトリック音声合成システムに用いられる音響モデルには隠れマルコフモデル (HMM) だけでなく，Deep Feed-forward Neural Network (DNN) や Recurrent Neural Network (RNN) がその高い性能から注目を集めている．これまで 1 時間から 20 時間程度のコーパスを用いたニューラルネットワークに基づく音声合成システムの構築・性能評価は報告されてるものの，さらに巨大なコーパスを用いた音声合成システムの構築はなされていない．本研究では 20 時間から 100 時間程度まで学習データ量を変更しつつ HMM・DNN・RNN 音声合成システムの構築を行い，評価実験を行った．

キーワード：音声合成，隠れマルコフモデル，Deep Neural Network

1. はじめに

統計的パラメトリック音声合成システムを実現する代表的な手法として，隠れマルコフモデル (Hidden Markov Model; HMM) に基づく枠組みが挙げられる [1]．HMM に基づく音声合成を用いることである程度高品質な音声の合成を実現できるが，決定木に基づくコンテキストクラスタリングにより学習データが分割されてしまうことや，出力分布として単純なガウス分布が状態単位で割り当てられるといった問題が存在する．近年では，このような問題に対してニューラルネットワークを用いることが検討されており，例えば，HMM 音響モデルとニューラルネットワークを組み合わせる手法 [2] や，HMM 音響モデルの枠組み全体をニューラルネットワークに置き換える手法 [3], [4], [5] が提案されている．ニューラルネットワークに基づく音声合成システムは高い性能を持つことが報告されている．

近年その高い性能からニューラルネットワークは注目を集めているが，ニューラルネットワークに基づく音声合成システムは 1990 年代まで遡ることができる [6], [7], [8]．ニューラルネットワークに基づくアプローチが再注目されている理由には，効果的な初期化手法 [9] が提案されたことや計算機環境の改善 (GPU の利用) 等による学習プロセスの効率化が要因として挙げられる．また，深層構造を持つニューラルネットワークを学習するために必要となる，大量のデータを利用可能になったことも要因として挙げられる．音声合成と同様に音響モデルの学習が行われる音声認識分野においては，学習データ量の違いによる単語認識率への影響が報告されている．Amodei らは学習データを増加させることでニューラルネットワークに基づく音声認識器による単語認識率は減少していき，全学習データ 12,000 時間を用いた際に最も性能が良いことを報告している [10]．また，HMM に基づく枠組みと比較してニューラルネットワークに基づく音声認識システムは，巨大学習データを有効に活用可能であるとも報告されている [11]．

音声合成分野ではニューラルネットワーク構築に 1 時間から 20 時間程度の学習データが用いられており [5]，音声認識器構築と比較すると小規模のデータサイズとなっている．これは音声認識では不特定話者を対象とし，また，ノイズに頑健な音響モデルが求められる一方で，音声合成システムの構築では特定話者データを用い，話者やノイズによる音響的特徴の変動を考慮しない場合が多いという点が理由として挙げられる．また，音声合成に用いられるコーパスは音素や韻律情報をバランス良く網羅するように設計する必要があり，適切な環境において音声データの収録を

[1] 総合研究大学院大学
SOKENDAI, Chiyoda, Tokyo 101–8430, Japan
[2] 国立情報学研究所
National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
a) wangxin@nii.ac.jp
b) takaki@nii.ac.jp
c) jyamagis@nii.ac.jp

行う必要がある．さらには，韻律情報のアノテーションといった手作業も必要になる場合がある．このように音声合成システム構築向けの巨大音声コーパスの用意には莫大なコストがかかってしまうが，小規模なデータであっても丁寧に設計されていれば，高品質なニューラルネットワークに基づく音声合成システムの構築が可能である．

その一方で，音声合成システムの構築向けに設計された巨大特定話者コーパスを用いることによる，合成音声の品質への影響を調査することも非常に興味深い．音声認識では学習データ量を増やすことによる性能への効果は徐々に小さくなっていくと報告されている [12], [13] が，音声合成ではスペクトルだけでなく F0 においても同様の学習データ量の違いによる予測精度への影響を見る必要がある．本論文では，20 時間から 100 時間まで学習データ量を変更し HMM，Deep Feed-forward Neural Network (DNN)，Recurrent Neural Network (RNN) に基づく音響モデルを用い音声合成システムをそれぞれ構築を行い，性能の比較を行った．

2. コーパス

実験では ATR において開発された波形接続型音声合成システム Ximera [14], [15] に用いられた，特定話者日本語音声コーパスを使用した．このコーパスには男性話者 M007 と女性話者 F009 のデータが含まれる．男性話者，女性話者の音声データはそれぞれ約 110 時間，約 50 時間である．男性話者，女性話者の音声はそれぞれ 973 日間 (内 181 日収録)，307 日間 (内 95 日収録) をかけ収録を行った．コーパスの収録文には新聞，小説，旅行対話文等が含まれる．また，手作業により音声データとテキストの対応付けが行われている．収録は同一のマイクロフォン，防音室を用いて行われ，24bit，48kHz サンプリングの音声が収録されている．日本において利用可能な最大級の音声合成向けコーパスと言える．

3. 特徴量抽出

ニューラルネットワークに基づく音声合成システム構築には，60 次メルケプストラム係数 (MGC)，log F0 (F0)，無声/有声パラメータ (U/V)，25 次非周期成分を用いた．これら音響特徴量は F0 適応窓を用いて STRAIGHT により各フレーム毎に抽出を行った [16]．フレームシフトは 5 ms である．また，無声/有声パラメータ以外のパラメータについてはその動的特徴量 (Δ，Δ^2) も用いている．RNN 音響モデルでは動的特徴量は必要でないとも考えられるが，本実験では比較のため他のシステムと同一の音響特徴量を用いることとした．ニューラルネットワークに基づく音響モデル構築に使用した音響特徴量は合計 259 次元となる．HMM に基づく音声合成システム構築には同様のメルケプストラム，非周期成分に加え，無声/有声を含む不連続な log F0 を用いており，音響特徴量は合計 258 次元となる．

ニューラルネットワークの入力として用いられる言語特徴量は音声合成システム Open JTalk [17] のテキスト解析部を用いて得た．Open JTalk のテキスト解析部では Mecab [18] により形態素，品詞解析や音素への変換等のテキスト解析が行われる．テキスト解析結果は 389 次元の言語特徴量に変換され，ニューラルネットワークに基づく音声合成システムの入力として用いられる．言語特徴量に含まれる音素継続長は HMM により推定を行った．HMM 音声合成システムには同様の情報を持つコンテキストラベルを入力として用いた．

4. 実験

4.1 実験条件

HMM に基づく音声合成システム (以下，HMM 音声合成システムと呼ぶ) の構築には HTS Toolkit Version2.3 を用いた．HMM は 5 状態のスキップなし left-to-right の隠れセミマルコフモデルとした．基本周波数パラメータは多空間分布 HMM (MSD-HMM) によりモデル化を行った [19]．まず，モノフォン HMM の学習を行い，コンテキスト依存 HMM への変換を行った．その後，頑健なモデルパラメータの推定のため，決定木に基づくコンテキストクラスタリングの適用を行った．

DNN に基づく音響モデルを用いた音声合成システム (以下，DNN 音声合成システムと呼ぶ) では隠れ層数が 4，ユニット数が入力層に近い層から 1024，512，512，512 のネットワークを使用した．RNN に基づく音響モデルを用いた音声合成システム (以下，DBLSTM-RNN 音声合成システムと呼ぶ) では隠れ層として Feed-forward 層 2 つと Bi-directional Recurrent 層 2 つを用いた．Feed-forward 層は 512 ユニットとし，Bi-directional Recurrent 層は 256 ユニットの Long-short Term Memory とした．本実検で用いた Deep Bidirectional Long Short-term Memory Recurrent Neural Network (DBLSTM-RNN) は文献 [4] にならい構造を決定した．

DNN と DBLSTM-RNN の学習には CURRENNT Toolkit [20] を用いた．全てのニューラルネットワークはランダムに初期化した後，確率的勾配法により学習を行った．DBLSTM-RNN 音声合成システムでは学習時間短縮のため，20 発話毎のパラレルセンテンス学習を用いた [20]．DNN 音声合成システムではこのパラレルセンテンス学習は用いていない．なお CURRENNT Toolkit は主として RNN の学習のために設計されており，各発話がミニバッチ一つとして扱われる．ラーニングレートには DNN と DBLSTM-RNN に基づく音声合成システム両方の学習において $4e-06$ を用いた．

音声合成時には HMM, DNN, DBLSTM-RMM より予測された Δ，Δ^2 を含むパラメータ系列に対して，Maximum

Likelihood Parameter Generation (MLPG) アルゴリズムを用い静的特徴量を得た [21]．DNN，DBLSTM-RNN 音声合成システムにおいて MLPG アルゴリズムを適用する際には，全学習データより各次元毎に計算したグローバルな分散を使用した．本実検では自然音声から得た音素アライメントをテストデータに利用した．また，全システムにおいて予測されたメルケプストラムには信号処理に基づくポストフィルタを適用した [22]．音声波形の生成は予測された音響特徴量から STRAIGHT を用い行った．

男性話者 M007 コーパス，女性話者 F009 コーパスにおいてそれぞれランダムに選択された 500 文と 260 文を評価・テストデータとして用い，残りのデータは学習データとして用いた．

4.2 客観評価結果

図 1，2 に M007 コーパスと F009 コーパスを用いて構築された音声合成システムの客観評価結果をそれぞれ示す．図中左から F0 相関係数，F0 RMSE，無声/有声エラー，メルケプストラム RMSE を示している．図 1，2 より，どの評価においても DBLSTM-RNN 音声合成システムの性能が最も高く，次いで DNN 音声合成システム，HMM 音声合成システムの性能順となっている．ここから音響モデリング手法による性能差が大きいことがわかる．例えば，男性話者の約 100 時間，約 20 時間のデータを用いてそれぞれ構築された HMM 音声合成システムと DNN 音声合成システムの F0 相関係数の結果は同程度である．また，約 100 時間，約 20 時間のデータを用いてそれぞれ構築された DNN 音声合成システムと DBLSTM-RNN 音声合成システムの F0 相関係数の結果も同程度である．DBLSTM-RNN 音声合成システムは，より多くの学習データを用いた際に見られる性能改善が大きい．

図 1，2 から，より多量のデータを用いる事による音声合成システムの性能向上を確認できるが，メルケプストラムと F0 それぞれの結果に異なる傾向も見られる．メルケプストラム RMSE はデータ量が増えるにつれて減少はしているが，改善は小さくなり収束傾向が見られる．一方で，DNN 音声合成システムの結果を除き，HMM 音声合成システムと DBLSTM-RNN 音声合成システムの F0 相関係数，F0 RMSE は学習データ量が増加するにつれ改善し続けている．また，M007 コーパス，F009 コーパス両方において HMM 音声合成システムと DBLSTM-RNN 音声合成システムの F0 RMSE の改善は，学習データが増加するとより大きくなっていることがわかる．

図 3 に M007 コーパスをから構築された音声合成システムにより予測された F0 軌跡の例を示す．図 3a には学習データ量の異なる DBLSTM-RNN 音声合成システムにより予測された F0 軌跡を，図 3b には全学習データを用いた HMM，DNN，DBLSTM-RNN 音声合成システムにより予測された F0 軌跡を示す．図には自然音声から抽出された F0 軌跡も示している．図 3a より全学習データを用いて構築された DBLSTM-RNN 音声合成システムが，250 から 350 のフレーム間における有声部において他のシステムよりも高い精度で F0 の予測を行っていることがわかる．しかし，図 3a 中の他のフレームにおいて多量の学習データを用いたことによる改善は見られない．この例では，学習データ量を増加させたことにより F0 軌跡が部分的に改善されていることがわかる．また，図 3b より HMM 音声合成システムと比較して DNN 音声合成システムから予測された F0 軌跡は，自然音声から抽出された F0 軌跡から大きく外れている部分があることがわかる．図 1，2 においても見られるように，DNN 音声合成システムでは F0 モデリングが適切に行われていない．他の手法と比較して，DNN では F0 軌跡の時間的変動を適切に捉えることができていない可能性がある．

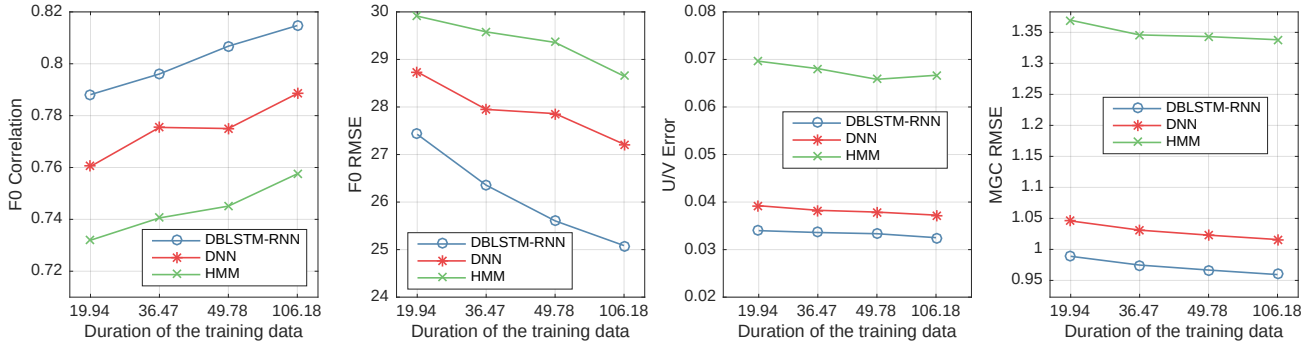
また，テストデータは違うものの M007 コーパス，F009 コーパス間での結果の比較も興味深い．M007 コーパス，F009 コーパスで構築された音声合成システムの F0 相関係数の結果を比較すると，より大きな M007 コーパスよりも F009 コーパスの音声合成システムが良い結果となっている．また，コーパス間の音声合成システムを用いた主観評価実験は行ってはいないが，筆者らは F009 コーパスにより構築された音声合成システムの合成音声の方が M007 の合成音声よりも高品質に感じた．より長期間に及ぶ音声収録時における男性話者 M007 の身体的，精神的コンディションの違いによる影響が考えられる．

4.3 主観評価結果

主観評価実験には M007 コーパス，F009 コーパスを用いて構築された音声合成システムをそれぞれ MUSHRA 法を用い評価した．自然音声を隠れアンカーとして使用した．被験者数は 7 人の日本語ネイティブ話者である．各被験者は被験者ごとにテスト文からランダムに選ばれた 15 文章，または，25 文章を比較した．聴取試験はヘッドホンを用いて静かな部屋で行った．

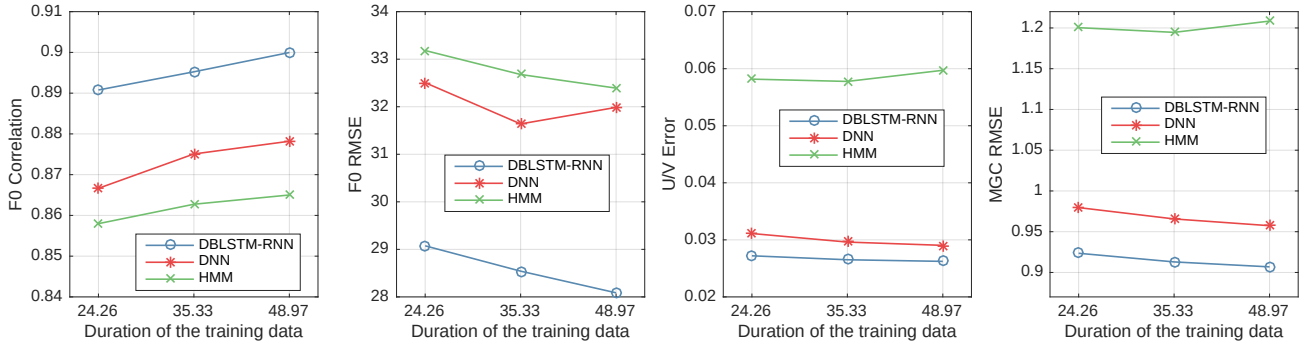
図 4a，4b にそれぞれ M007 コーパス，F009 コーパスを用いて構築された音声合成システムの主観評価結果を示す．図 4a では客観評価において用いたシステムから HMM，DNN，DBLSTM-RNN 音声合成システムからそれぞれ 3 種類を選択しており，図 4b では全てのシステムが用いられている．

まず，両コーパスにおいて客観評価結果と同様に DBLSTM-RNN 音声合成システムの性能が最も高く，次いで DNN 音声合成システム，HMM 音声合成システムの性能順になっている．M007 コーパスにおいて約 100 時間データを用いた DNN 音声合成システムと約 20 時間データを用いた DBLSTM-RNN 音声合成システムは同程度の



(a) F0 Correlation (b) F0 RMSE (c) U/V Error (d) MGC RMSE

図 1: Performance of the DBLSTM-RNN, DNN and HMM on the M007 corpus.



(a) F0 Correlation (b) F0 RMSE (c) U/V Error (d) MGC RMSE

図 2: Performance of the DBLSTM-RNN, DNN and HMM on the F009 corpus.

性能となっており，主観評価実検においても音響モデリング手法の違いが大きな性能差として表れている．

また，客観評価結果とは異なり，学習データ量を増やしたことによる HMM 音声合成システムの性能改善は小さく (M007 コーパス)，もしくは，全く見られない (F009 コーパス)．DNN，DBLSTM-RNN 音声合成では学習データ量の増加による改善はある程度見て取れる (DNN，DBLSTM-RNN 音声合成それぞれにおける M007 コーパス約 20 時間と約 100 時データ間での比較) が，例えば約 50 時間と約 100 時間データによる男性話者 DBLSTM-RNN 音声合成システムのように，DBLSTM-RNN 音声合成システムにおいて学習データ量を増加させても主観評価結果が同程度のシステムもある．このような結果になったのは，自動ラベリングを用いているためラベリングエラーが増大してしまったことや，学習データを増加させたとしてもメルケプストラムの改善は収束傾向にあることが要因として挙げられるが，F0 相関係数，F0 RMSE の大幅な改善が合成音声の知覚上の改善に繋がらなかったとも考えられる．

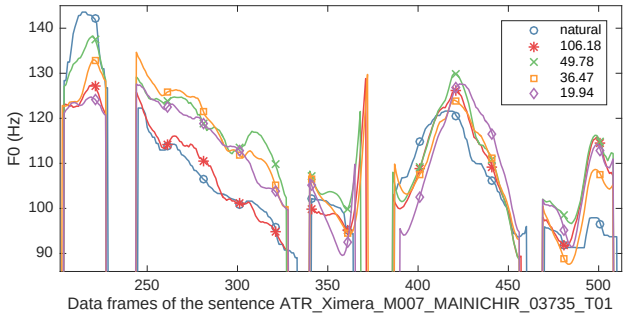
5. おわりに

本論文では学習データ量の違いによる HMM，DNN，RNN に基づく音声合成システムの性能への影響を客観評価，主観評価により調査を行った．音声合成システムは特定男性話者 100 時間，特定女性話者 50 時間のデータを用い構築を行った．どのシステムにおいても学習データ量を増加させることによりメルケプストラムの予測誤差には収束傾向が見られたが，F0 の予測において HMM と RNN に基づく音声合成システムは全てのデータを使用するまで，学習データの増加による性能改善が見られた．

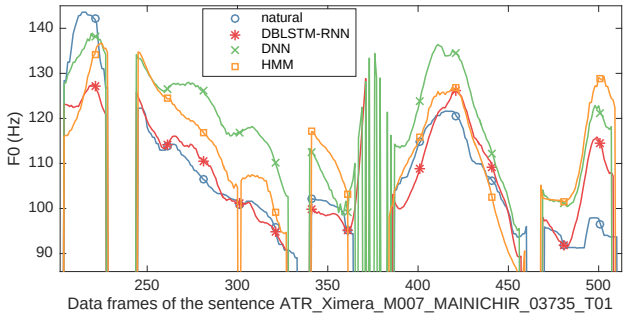
その一方で，主観評価実験では学習データ量を増やしたことによる HMM 音声合成システムの性能改善は小さく，また，DBLSTM-RNN 音声合成システムにおいても学習データ量の増加によりある程度の性能改善は見られたものの改善は限定的であった．

巨大データを用いた音声合成システム構築における今後の課題として，まず，Highway Network [23] を利用した，より多層のネットワークの構築が挙げられる．より多層のネットワークを用いることで，メルケプストラムの予測においても巨大データを用いることによる性能改善が得られる可能性があると考えている．その他，人工的なデータを用い学習データを増やすことによる音声合成システム構築も課題として挙げられる．

また，巨大データの利用に適した F0 モデリング手法の検討も興味深い．文献 [24] にて報告されているように，F0

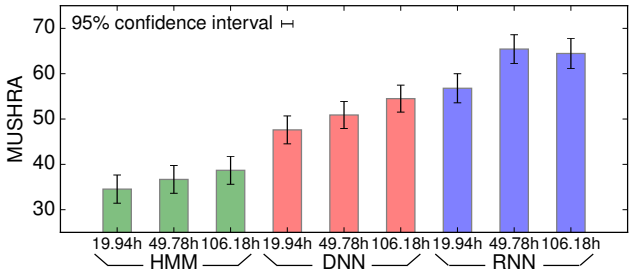


(a) The DBLSTM-RNN systems trained with different amount of training data are compared.

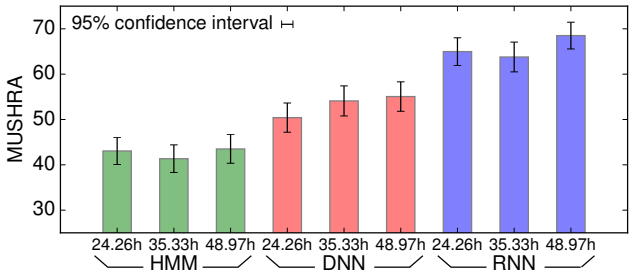


(b) The DBLSTM-RNN, DNN and HMM systems trained using the full M007 data are compared.

図 3: Samples of synthetic F0 trajectory on the M007 voice.



(a) M007 Corpus.



(b) F009 Corpus.

図 4: Subjective results (MUSHRA).

軌跡をある程度直線に置き換えたとしても抑揚に関して知覚上の劣化は少ない．ニューラルネットワークの学習において，正確に F0 軌跡を表現するように評価関数を設計することは適切でない可能性がある．また，平均二乗誤差を評価関数として用いて学習されたニューラルネットワークに基づくシステムは，学習データの特徴量の平均値を予測する傾向にある [25]．音声合成において F0 特徴量の分布に多峰性があれば，平均 F0 パターンが予測されることで良い客観評価結果が得られるであろうが，音声サンプルとしては好まれないことが予想される．より詳細な合成音声サンプル，及び，ニューラルネットワークの解析が必要である．

謝辞 本研究の一部は MEXT 科研費 JP16K16096，電気通信普及財団，NAVER Lab. の助成を受けた．

参考文献

[1] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura: Speech synthesis based on hidden Markov models, *Proceedings of the IEEE*, Vol. 101, No. 5, pp. 1234–1252 (2013).

[2] Z.-H. Ling, L. Deng, and D. Yu: Modeling Spectral Envelopes Using Restricted Boltzmann Machines and Deep Belief Networks for Statistical Parametric Speech Synthesis, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, pp. 2129–2139 (2013).

[3] Heiga Zen, Alan Senior, and Martin Schuster: Statistical parametric speech synthesis using deep neural networks, *ICASSP-2013*, pp. 7962–7966 (2013).

[4] Y. Fan, Y. Qian, F. Xie, and F. K. Soong: TTS Synthesis with Bidirectional LSTM Based Recurrent Neural Networks, *INTERSPEECH-2014*, pp. 1964–1968 (2014).

[5] Shiyin Kang and Helen M. Meng: Statistical parametric speech synthesis using weighted multi-distribution deep belief network, *INTERSPEECH-2014*, pp. 1959–1963 (2014).

[6] Christine Tuerk and Tony Robinson: Speech synthesis using artificial neural networks trained on cepstral coefficients., *EUROSPEECH*, pp. 1713–1716 (1993).

[7] Orhan Karaali, Gerald Corrigan, and Ira Gerson: Speech Synthesis with Neural Networks, *Proc. of World Congress on Neural Networks*, pp. 45–50 (1996).

[8] Sin-Horng Chen, Shaw-Hwa Hwang, and Yih-Ru Wang: An RNN-based prosodic information synthesizer for Mandarin text-to-speech, *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 3, pp. 226–239 (1998).

[9] G. E. Hinton and R. Salakhutdinov: Reducing the dimensionality of data with neural networks, *Science 28*, Vol. 313, No. 5786, pp. 504–507 (2006).

[10] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, et al.: Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, *arXiv preprint arXiv:1512.02595* (2015).

[11] Xuedong Huang, James Baker, and Raj Reddy: A Historical Perspective of Speech Recognition, *Commun. ACM*, Vol. 57, No. 1, pp. 94–103 (online), DOI: 10.1145/2500887 (2014).

[12] Kai Wei, Yuzong Liu, Katrin Kirchhoff, Christopher Bartels, and Jeff Bilmes: Submodular subset selection for large-scale speech training data, *ICASSP-2014*, IEEE, pp. 3311–3315 (2014).

[13] Yi Wu, Rong Zhang, and Alexander Rudnicky: Data selection for speech recognition, *ASRU-2007*, IEEE, pp. 562–565 (2007).
[14] Hisashi Kawai, Tomoki Toda, Jinfu Ni, Minoru Tsuzaki, and Keiichi Tokuda: XIMERA: A new TTS from ATR based on corpus-based technologies, *Fifth ISCA Workshop on Speech Synthesis* (2004).
[15] H Kawai, T Toda, J Yamagishi, T Hirai, J Ni, N Nishizawa, M Tsuzaki, and K Tokuda: XIMERA: A concatenative speech synthesis system with large scale corpora, *IEICE Trans. Inf. Syst.(Japanese Edition)*, pp. 2688–2698 (2006).
[16] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds, *Speech Communication*, Vol. 27, pp. 187–207 (1999).
[17] The HTS Working Group: The Japanese TTS System "Open JTalk" (2015).
[18] Taku Kudo: MeCab: Yet Another Part-of-Speech and Morphological Analyzer.
[19] Keiichi Tokuda, Takashi Masuko, Noboru Miyazaki, and Takao Kobayashi: Multi-space probability distribution HMM, *IEICE TRANSACTIONS on Information and Systems*, Vol. 85, No. 3, pp. 455–464 (2002).
[20] Felix Weninger, Johannes Bergmann, and Björn Schuller: Introducing CURRENT: The Munich open-source CUDA recurrent neural network toolkit, *The Journal of Machine Learning Research*, Vol. 16, No. 1, pp. 547–551 (2015).
[21] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura: Speech parameter generation algorithms for HMM-based speech synthesis, *Proceedings of ICASSP 2000*, pp. 936–939 (2000).
[22] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura: Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis, *IEICE*, Vol. J87-D-II, No. 8, pp. 1565–1571 (2004).
[23] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber: Highway Networks, *CoRR*, Vol. abs/1505.00387 (online), available from ⟨http://arxiv.org/abs/1505.00387⟩ (2015).
[24] Christophe d'Alessandro and Piet Mertens: Automatic pitch contour stylization using a model of tonal perception, *Computer Speech & Language*, Vol. 9, No. 3, pp. 257 – 288 (1995).
[25] Christopher M Bishop: Mixture density networks, Technical report, Aston University (2004).