Blizzard Challenge 2017 のための NITech テキスト音声合成システム *

◎沢田慶, 橋本佳, 大浦圭一郎, 南角吉彦, 徳田恵一 (名工大)

1 はじめに

本稿では，Blizzard Challenge 2017 のために開発した NITech テキスト音声合成 (text-to-speech; TTS) システムの構成と評価結果について述べる．Blizzard Challenge 2017 では，表現豊かな音声が数多く含まれる児童書のオーディオブックが配布され，参加者はこれを用いて TTS システムを構築する．我々は，高品質な統計的パラメトリック音声合成 (statistical parametric speech synthesis; SPSS) システムを実現するために，学習コーパスの自動構築法の検討と言語特徴量の設計を行った．音響モデルには，系列内変動を考慮したトラジェクトリ学習による混合密度ネットワーク (mixture density network; MDN) を用いた．大規模主観評価実験の結果，NITech TTS システムは高い自然性と，提出された TTS システムの中で最も高い明瞭性を示した．

2 Blizzard Challenge 2017

近年，コーパスベースの手法による TTS システムが数多く提案されている．しかし，TTS システムを比較・評価するためには，同一の学習コーパスでシステムを構築する必要がある．また，TTS システムの評価には受聴試験による主観評価が広く用いられているが，別々に実施された受聴試験の結果を直接比較することは適切ではない．そこで，同一の学習コーパスと受聴試験により，TTS システムの比較・評価を行う Blizzard Challenge が開催されている [1].

Blizzard Challenge 2017 では，英語の児童書のオーディオブック 56 冊 (約 7 時間) が学習コーパスとして配布された [2]．オーディオブックは，書籍を朗読した音声を収録したコンテンツである．オーディオブックは，音声とテキストの対を大量に収集することができることや，同一話者による比較的クリーンかつ同一環境で収録された音声が多く存在することから，TTS システムのための学習コーパスとして注目を集めている．しかし，オーディオブックの音声とテキストには，読み間違いやテキストに存在しない文 (書籍の説明や擬音語等) を追加情報として収録したことによる不一致が存在する．さらに，児童書のオーディオブックは，複数のキャラクターが登場することや子供を楽しませるために表現豊かに読み上げていることから様々な発話スタイルの音声が含まれている．そのため，表現豊かな音声合成を実現するための学習コーパスとして適している．参加者は，このようなデータを用いて児童に読み上げるのに最適な TTS システムの構築を目指す．

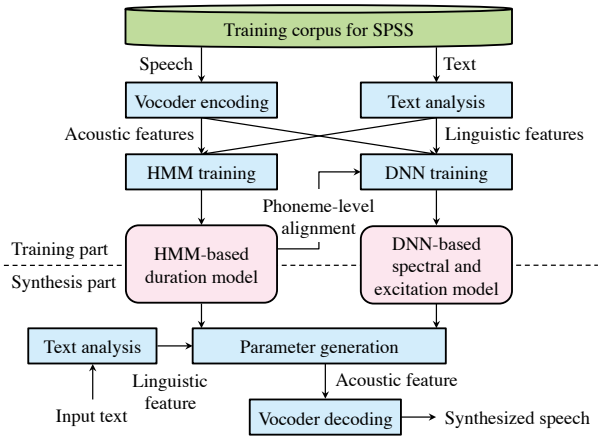

Fig. 1 Overview of the NITech TTS system

3 NITech テキスト音声合成システム

NITech TTS システムの概要を Fig. 1 に示す．学習部では，テキスト解析によりテキストから言語特徴量を，ボコーダにより音声から音響特徴量を推定する．次に，言語特徴量と音響特徴量から隠れマルコフモデル (hidden Markov model; HMM) に基づく音声合成システム (HMM 音声合成) を構築し，音響特徴量系列に対する言語特徴量系列の音素アライメントを推定する．最後に，音素アライメントの結果に基づいたフレーム単位の言語特徴量と音響特徴量の対からディープニューラルネットワーク (deep neural network; DNN) に基づく音声合成システム (DNN 音声合成) を構築する．合成部では，入力テキストの言語特徴量から，HMM 音声合成により音素継続長を，DNN 音声合成によりフレーム単位の音響特徴量を推定する．そして，ボコーダを用いて音声波形を生成する．

NITech TTS システムでは，高品質な音声を合成するために 1) 音声とテキストの不一致に対応するための音声認識を用いた学習コーパスの自動構築法，2) 児童書のオーディオブックに適する言語特徴量の設計，3) 系列内変動を考慮したトラジェクトリ学習による MDN の 3 点について検討した．

3.1 オーディオブックを用いた学習コーパスの自動構築

オーディオブックの音声とテキスト (ブックテキスト) には，読み間違いやテキストに存在しない文の収録による不一致が存在する．この不一致は音響モデルの学習に悪影響を与えるため，学習コーパスのテキストには音声と一致するテキスト (正解テキスト) を用いることが望ましい．しかし，大量の音声の正

* Overview of the NITech text-to-speech system for the Blizzard Challenge 2017.
by SAWADA, Kei, HASHIMOTO, Kei, OURA, Keiichiro, NANKAKU, Yoshihiko, and TOKUDA, Keiichi
(Nagoya Institute of Technology; NITech)

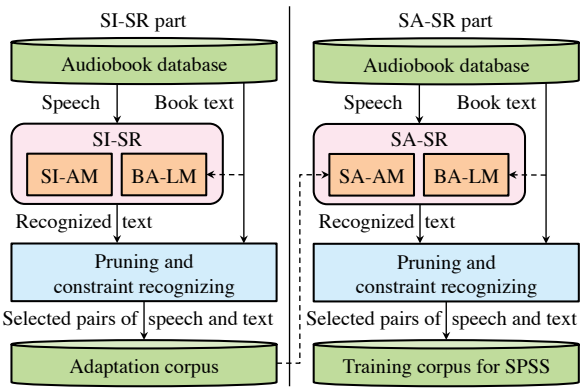

Fig. 2 Overview of training corpus construction (SI: speaker independent, SA: speaker adapted, BA: book adapted, SR: speech recognizer, AM: acoustic model, LM: language model)

解テキストを人手により得ることは高コストである．そこで，音声認識により音声の発話内容に対応するテキスト (認識テキスト) を推定する．Fig. 2 にオーディオブックを用いた学習コーパス自動構築の流れを示す [3].

正解テキストとブックテキストの大部分は一致していると考えられる．そのため，ブックテキストを学習コーパスとする言語モデルはオーディオブックの音声認識に非常に有用である．そこでブック毎に，大量のテキストとブックテキストを学習コーパスとしたブック適応言語モデルを用いることで，音声認識の性能改善を図る．

Fig. 2 の不特定話者音声認識部 (SI-SR part) では，不特定話者音響モデルとブック適応言語モデルを用いて認識テキストを推定する．ここで，ブックテキストと認識テキストの単語一致度が低い場合には，その認識テキストは正解テキストと大きく異なる可能性がある．そこで，単語一致度が高い音声のみを適応コーパスとして話者適応音響モデルを構築する．話者適応音声認識部 (SA-SR part) では，話者適応音響モデルとブック適応言語モデルを用いて音声認識を行い，単語一致度が高い音声を SPSS の学習コーパスとする．

さらに，ブックテキストと認識テキストから正解テキストを推定する手法について検討する．ブックテキストと認識テキストから単語ネットワークを構築し，この単語ネットワークに制限した音声認識により，学習コーパスとして利用するテキスト (制限認識テキスト) を推定する．制限認識テキストは，認識エラーを削減しつつ，追加情報に対応するテキストを推定することが可能であることから，正解テキストに近いテキストであることが期待される．

3.2 オーディオブックのための言語特徴量の設計

児童書のオーディオブックには，多様な発話スタイルの音声が収録されているため，表現豊かな音声合成を実現するための学習コーパスとして適している．しかし，これまでに用いられてきた SPSS の言語特徴量は，オーディオブックを想定して設計されていない．そこで，児童書のオーディオブックを学習コーパスとして音響モデルを学習することを想定し，言語特徴量を設計する．

従来の言語特徴量には，文単位までの情報が用いられてきた．しかし，オーディオブックは前後の文章が意味的に関係するため，文単位までの言語特徴量では不十分であると考えられる．本稿で学習コーパスとする児童書はページ単位で場面が切り替わるため，新たにページ単位の言語特徴量を導入する．また，文の構造は韻律に影響するため，文の構造を捉えた言語特徴量は音響モデルの学習に有用である [4]．そこで文の構造を捉えるために，構文解析と係り受け解析から得られる情報を言語特徴量として利用する．

児童書のオーディオブックは，様々な発話スタイルを含んでいる．特に，会話文や感嘆文は強調や感情表現が多く含まれており，これらの情報を言語特徴量によって表現することで，より適切に発話スタイルをモデル化することが可能となる．そこで，引用符と文の種類に関する言語特徴量を導入し，会話文や感嘆文の発話スタイルを明確に区別する．

さらに，単語・句コードを言語特徴量に追加する．句コードは，学習コーパスの句毎に割り振られた一意の値である．同一のテキストが異なる発話スタイルで収録されていたとしても，句コードが異なることで句毎に発話スタイルを区別した学習ができる．合成時には，適切な句コードを設定することで様々な発話スタイルを表現することが可能となる．しかし，テスト文毎に句コードを人手で与えることはコストが高い．そこで本稿では，テスト文の句コードを，学習コーパス中の句コードから自動的に選択する枠組みを検討する．句コードを自動選択するために，自然言語処理の分野で提案されている文章をベクトル化する技術である doc2vec [5] を用いる．文章をベクトル化することにより，文章間の類似度を計算することが可能となる．学習コーパスとテスト文の句を doc2vec によりベクトル化して類似度を計算し，学習コーパス中で最も類似度が高い句コードをテスト文の句コードとする．例えば，テスト文が怒っている内容の句の場合には，学習コーパス中の怒っている内容の句と類似度が高くなる．そして，類似度が最も高い句が怒っている発話スタイルであれば，テスト文も怒っている発話スタイルで合成することが可能となる．このようにして，テスト文から発話スタイルを自動選択することで，様々な発話スタイルの音声を合成することができる．単語コードについても同様に，word2vec [6] を用いて単語の類似度を計算し，類似度の高い単語コードをテスト文の単語コードとする．

3.3 系列内変動を考慮したトラジェクトリ学習を用いた混合密度ネットワーク

DNN 音声合成における，DNN の構造には MDN を用いる [7]．MDN では，言語特徴量から正規混合モデルのモデルパラメータを推定する DNN を学習

する．ただし，本稿では MDN の混合数は 1 とする．MDN の学習は，次式で表される評価関数 $\mathcal{L}$ を最大化する．

$$\mathcal{L} = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t \,|\, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \tag{1}$$

ここで，T は音響特徴量のフレーム数，$\boldsymbol{o}_t$ は時刻 t における静的・動的特徴量，$\mathcal{N}(\cdot \,|\, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ は平均 $\boldsymbol{\mu}_t = [\mu_{t,1}, \mu_{t,2}, \ldots, \mu_{t,D}]^\top$，共分散行列 $\boldsymbol{\Sigma}_t = \mathrm{diag}[\sigma^2_{t,1}, \sigma^2_{t,2}, \ldots, \sigma^2_{t,D}]$ の正規分布，D は音響特徴量の次元数を表す．MDN の出力となるパラメータは以下によって表される．

$$\mu_{t,d} = g_d^{(\boldsymbol{\mu})}(\boldsymbol{l}_t, \boldsymbol{\lambda}_{\mathrm{MDN}}) \tag{2}$$

$$\sigma_{t,d} = \exp(g_d^{(\boldsymbol{\sigma})}(\boldsymbol{l}_t, \boldsymbol{\lambda}_{\mathrm{MDN}})) \tag{3}$$

ただし，$g_d^{(\boldsymbol{\mu})}(\boldsymbol{l}_t, \boldsymbol{\lambda}_{\mathrm{MDN}})$ と $g_d^{(\boldsymbol{\sigma})}(\boldsymbol{l}_t, \boldsymbol{\lambda}_{\mathrm{MDN}})$ は言語特徴量 $\boldsymbol{l}_t$ と MDN のパラメータ $\boldsymbol{\lambda}_{\mathrm{MDN}}$ が与えられたもとでの出力層の活性化関数である．合成時には静的・動的特徴量の関係を考慮したパラメータ生成を行う．

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} \mathcal{N}(\boldsymbol{W}\boldsymbol{c} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{4}$$

ここで，$\boldsymbol{c}$ は静的特徴量系列を表し，$\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ は MDN の出力特徴量 $\boldsymbol{\mu}_t$ と $\boldsymbol{\Sigma}_t$ をそれぞれ連結したパラメータを表す．また，$\boldsymbol{W}$ は静的特徴量系列から静的・動的特徴量系列へと変換する窓行列である．

DNN 音声合成において，系列内変動を考慮したトラジェクトリ学習の有効性が報告されている [8]．本稿では，系列内変動を考慮したトラジェクトリ学習を MDN の学習に適用する．MDN の出力を系列内変動を考慮したトラジェクトリモデルのパラメータとしたとき，評価関数 $\mathcal{L}_{\mathrm{GVTrj}}$ は以下のように表される．

$$\mathcal{L}_{\mathrm{GVTrj}} = \mathcal{N}(\boldsymbol{c} \,|\, \bar{\boldsymbol{c}}, \boldsymbol{P}) \mathcal{N}(\boldsymbol{v}(\boldsymbol{c}) \,|\, \boldsymbol{v}(\bar{\boldsymbol{c}}), \boldsymbol{\Sigma}_{\mathrm{GV}})^{wT} \tag{5}$$

ここで，$\bar{\boldsymbol{c}}$ と $\boldsymbol{P}$ は次式とする．

$$\bar{\boldsymbol{c}} = \boldsymbol{P}\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \tag{6}$$

$$\boldsymbol{P} = \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\right)^{-1} \tag{7}$$

また，$\bar{\boldsymbol{c}}$ は式 (4) によって表されるパラメータ生成の出力 $\hat{\boldsymbol{c}}$ と一致する．そして，$\boldsymbol{v}(\boldsymbol{c})$ は静的特徴量系列 $\boldsymbol{c}$ の系列内変動，$\boldsymbol{\Sigma}_{\mathrm{GV}}$ は系列内変動の共分散行列，w は静的特徴量系列の尤度と系列内変動の尤度の関係を調整するパラメータ (GV 重み) である．Fig. 3 に MDN のための系列内変動を考慮したトラジェクトリ学習の概要を示す．合成時には，式 (5) の評価関数について最適な静的特徴量系列が生成されるが，この特徴量系列は式 (4) と完全に一致する．

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} \mathcal{N}(\boldsymbol{c} \,|\, \bar{\boldsymbol{c}}, \boldsymbol{P}) \mathcal{N}(\boldsymbol{v}(\boldsymbol{c}) \,|\, \boldsymbol{v}(\bar{\boldsymbol{c}}), \boldsymbol{\Sigma}_{\mathrm{GV}}) \tag{8}$$

このため，系列内変動を考慮した学習が行われた場合も，従来と同じ形で静的特徴量系列が推定される．

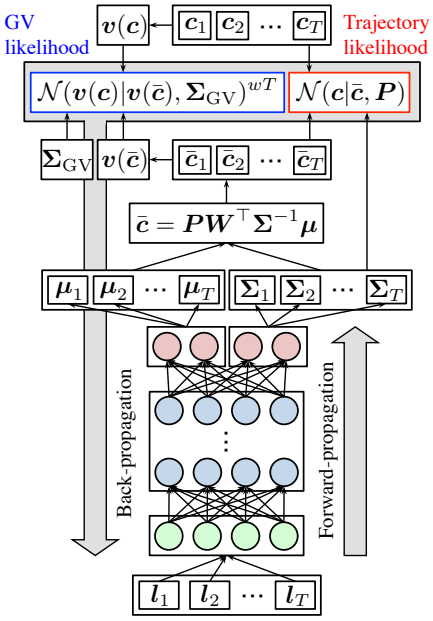

Fig. 3　Overview of trajectory training considering GV for MDN-based SPSS

4　Blizzard Challenge 2017 評価実験

本稿で述べた TTS システムを用いて，Blizzard Challenge 2017 に参加した．Blizzard Challenge 2017 では，同一話者により収録された英語の児童書 56 冊の音声とテキストが学習コーパスとして配布された．

4.1　実験条件

不特定話者音声認識器は，CMU Pronouncing Dictionary と WSJ0, WSJ1, TIMIT を用いて構築した．サンプリング周波数 16kHz の音声に対して，ハミング窓，フレーム長 25ms，フレームシフト 10ms とした 12 次元の MFCC とその 1, 2 次動的特徴量を音響特徴量とした．音響モデルは GMM-HMM，言語モデルは単語 tri-gram とし，それぞれ HTK と SRILM を用いて学習した．SPSS の学習コーパスとするための基準である単語一致度の閾値は 90%とし，全 1258 ページ中 921 ページを学習コーパスとして用いた．

SPSS のための言語特徴量は Festival，Stanford Parser, SyntaxNet, gensim を用いて推定した．言語特徴量は 925 次元のテキストから推定した特徴，10 次元の継続長特徴，150 次元の単語コード，600 次元の句コードから構成される計 1685 次元である．ここで，単語・句コードは固定次元の平均 0 分散 1 の正規分布に基づく乱数ベクトルとした．また，word2vec と doc2vec は事前学習されたモデルを利用した．音響特徴量には，サンプリング周波数 44.1kHz の音声に対してフレームシフト 5ms で推定した STRAIGHT メルケプストラム，F_0，非周期成分のメルケプストラムを用いた．RAPT, SWIPE', REAPER を用いて F_0 を抽出し，多数決の結果を最終的な F_0 とした．HMM 音声合成の音響モデルには，49 次のメルケプストラム，F_0，24 次の非周期成分のメプケプトラムとそれらの 1, 2 次動的特徴量を音響特徴量として学習した

Table 1　Evaluation results

System	Page domain							Sentence domain		SUS
	OI	PL	SP	ST	IN	EM	LE	NAT	SIM	WER
A	47 ± 9*	46 ± 10*	47 ± 9*	47 ± 10*	47 ± 10*	47 ± 10*	47 ± 9*	4.7 ± 0.6*	4.5 ± 0.8*	–
I	38 ± 10*	38 ± 10*	36 ± 11*	36 ± 12*	37 ± 11*	38 ± 11*	36 ± 11*	4.0 ± 0.9*	4.0 ± 0.9*	34
G	32 ± 10	31 ± 10	34 ± 11	33 ± 11	33 ± 10*	34 ± 11	32 ± 10	3.6 ± 0.9	3.1 ± 1.2	35
L	**31 ± 10**	**30 ± 10**	**31 ± 12**	**31 ± 12**	**31 ± 11**	**33 ± 11**	**31 ± 10**	**3.6 ± 0.9**	**3.0 ± 1.1**	**30**
E	31 ± 12	32 ± 12	27 ± 13*	27 ± 13*	28 ± 13*	32 ± 12	28 ± 12*	3.5 ± 1.0	3.9 ± 0.9*	38*
P	31 ± 11	31 ± 10	31 ± 11	32 ± 12	32 ± 11	34 ± 11	30 ± 10	3.4 ± 1.0	3.6 ± 1.0*	38*
B	27 ± 11*	27 ± 11*	25 ± 12*	26 ± 12*	27 ± 12*	30 ± 11*	25 ± 11*	3.3 ± 1.0*	3.7 ± 1.0*	42*
M	26 ± 10*	25 ± 10*	24 ± 11*	27 ± 11*	26 ± 11*	26 ± 11*	25 ± 10*	3.2 ± 0.9*	3.1 ± 1.0	33
K	26 ± 10*	27 ± 10*	26 ± 12*	25 ± 12*	26 ± 11*	28 ± 11*	25 ± 10*	3.1 ± 1.0*	2.9 ± 1.1	42*
Q	25 ± 10*	26 ± 10*	23 ± 11*	23 ± 12*	24 ± 11*	28 ± 11*	23 ± 10*	3.1 ± 1.1*	2.8 ± 1.1	44*
D	25 ± 10*	23 ± 10*	30 ± 12	28 ± 12*	26 ± 11*	24 ± 11*	26 ± 10*	2.7 ± 0.9*	2.5 ± 1.0*	30
H	24 ± 9*	24 ± 9*	25 ± 11*	25 ± 11*	23 ± 10*	21 ± 10*	24 ± 9*	2.7 ± 1.0*	2.1 ± 0.9*	39*
J	22 ± 9*	21 ± 9*	24 ± 11*	23 ± 11*	22 ± 10*	20 ± 11*	22 ± 10*	2.7 ± 1.0*	2.6 ± 1.0*	35
F	21 ± 10*	22 ± 10*	24 ± 11*	24 ± 11*	24 ± 11*	25 ± 11*	21 ± 10*	2.6 ± 1.0*	2.4 ± 1.0*	46*
C	16 ± 8*	15 ± 8*	21 ± 11*	19 ± 10*	18 ± 10*	18 ± 10*	17 ± 9*	1.7 ± 0.7*	1.9 ± 0.9*	38*
O	11 ± 7*	11 ± 7*	22 ± 12*	19 ± 12*	16 ± 11*	16 ± 10*	11 ± 7*	1.5 ± 0.7*	1.5 ± 0.7*	60*
N	8 ± 6*	8 ± 6*	16 ± 11*	13 ± 11*	12 ± 9*	11 ± 8*	7 ± 6*	1.1 ± 0.5*	1.2 ± 0.5*	85*

MSD-HMM を用いた．DNN 音声合成の音響モデルには，69 次のメルケプストラム，無声部を有声部で線形補間した F_0 と有声無声情報，34 次の非周期成分のメプケプトラムを音響特徴量として学習した MDN を用いた．MDN の学習のために，言語特徴量は 0–1 で正規化し，音響特徴量は平均 0 分散 1 の正規分布となるように正規化した．MDN の中間層は 3 層の 8000 ユニット，ドロップアウト率は 0.6，GV 重みは 0.001 である．後処理として，パワーに対してダイナミックレンジコンプレッサーをかけた．

4.2　評価結果

Blizzard Challenge 2017 の主催者により大規模主観評価実験が行われた．ページ単位の合成音声を評価するために 7 つの基準 (overall impression; OI, pleasantness; PL, speech pause; SP, stress; ST, intonation; IN, emotion; EM) の 60 段階 MOS 試験が，文単位の合成音声を評価するために 2 つの基準 (naturalness; NAT, similarity; SIM) の 5 段階 MOS 試験が実施された．また，合成音声の明瞭性を評価するために無意味文 (SUS) の書き取りテストが行われた．表 1 に各基準の MOS と書き取りテストの単語誤り率 (WER) を示す．表 1 において，システム A は自然発話，システム B, C, D はベンチマークシステム，システム E から Q は参加者のシステムを表し，NITech TTS システムは **L** である．システムの順番は，NAT の MOS を基準にソートされている．また，* はシステム **L** と有意な差があるシステムを表す．

表 1 より，システム **L** は，ページ単位の評価で高い MOS を獲得した．さらに，NAT の MOS も高いことから，NITech TTS システムは自然性の高い音声を合成できたと言える．ただし，SIM については上位の MOS を得ることができなかったため，話者類似性については今後改善する必要がある．また，参加システムの中で最も低い WER を達成したことから，明瞭性が高い合成音声であることが示された．

5　むすび

本稿では，Blizzard Challenge 2017 のための NITech TTS システムの構成と評価結果を述べた．オーディオブックを用いた学習コーパス自動構築法と SPSS のための言語特徴量の設計を行った．音響モデルには，系列内変動を考慮したトラジェクトリ学習による MDN を用いた．大規模主観評価実験の結果，NITech TTS システムは高い自然性と明瞭性を示した．今後の課題として，ボコーダの利用による音声品質の劣化を緩和するために波形を直接モデル化する手法 (WaveNet [9] など) の導入などが挙げられる．

謝辞　本研究開発の一部は総務省 SCOPE (受付番号 162106106) の委託を受けたものです．

参考文献

[1] A.W. Black *et al.*, “The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets,” Interspeech 2005, pp. 77–80, 2015.
[2] Blizzard Challenge 2017, http://www.synsig.org/index.php/Blizzard_Challenge_2017.
[3] K. Sawada *et al.*, “The NITech text-to-speech system for the Blizzard Challenge 2016,” Blizzard Challenge 2016 Workshop, 2016.
[4] R. Dall *et al.*, “Redefining the linguistic context feature set for HMM and DNN TTS through position and parsing,” Interspeech 2016, pp. 2851–2855, 2016.
[5] Q. Le *et al.*, “Distributed representations of sentences and documents,” ICML 2014, 2014.
[6] T. Mikolov *et al.*, “Efficient estimation of word representations in vector space,” arXiv:1301.3781, 2013.
[7] H. Zen *et al.*, “Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis,” ICASSP 2014, pp. 3844–3848, 2014.
[8] K. Hashimoto *et al.*, “Trajectory training considering global variance for speech synthesis based on neural networks,” ICASSP 2016, pp. 5600–5604, 2016.
[9] A. van den Oord *et al.*, “WaveNet: A generative model for raw audio,” arXiv:1609.03499, 2016.