社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

DNN に基づく話し言葉音声合成における追加コンテキストの効果

山下　優樹† 　郡山　知樹†† 　齋藤　佑樹†† 　高道慎之介†† 　井島　勇祐†††
増村　　亮††† 猿渡　　洋††

† 東京大学工学部　〒113–8656 東京都文京区本郷 7 丁目 3–1
†† 東京大学大学院情報理工学系研究科　〒113–8656 東京都文京区本郷 7 丁目 3–1
††† NTT メディアインテリジェンス研究所　〒239–0847 神奈川県横須賀市光の丘 1–1
E-mail: †yukiyama913@g.ecc.u-tokyo.ac.jp, ††tomoki_koriyama@ipc.i.u-tokyo.ac.jp

あらまし　ディープニューラルネットワーク (DNN) に基づく音声合成では, パラ言語, 非言語情報を追加することで, 読み上げ音声よりも自発性の高い音声を再現できる. 本稿では, 日本語話し言葉コーパス (CSJ) に付与されている豊富なアノテーションを利用して, DNN に基づく話し言葉音声合成におけるパラ言語的, 非言語的特徴量の効果を評価する. 実験では, パラ言語的情報を付加することで, より高い再現性で話し言葉音声を合成できることを示す.
キーワード　音声合成, コンテキスト, 話し言葉音声, アノテーション, ディープニューラルネットワーク

The Effectiveness of Additional Context in DNN-based Spontaneous Speech Synthesis

Yuki YAMASHITA†, Tomoki KORIYAMA††, Yuki SAITO††, Shinnosuke TAKAMICHI††, Yusuke IJIMA†††, Ryo MASUMURA†††, and Hiroshi SARUWATARI††

† Faculty of Engineering, The University of Tokyo　7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–8656 Japan
†† Graduate School of Information Science and Technology, The University of Tokyo　7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–8656 Japan
††† NTT Media Intelligence Laboratories, NTT Corporation　1–1 Hikarinooka, Yokosuka-shi, Kanagawa, 239–0847 Japan
E-mail: †yukiyama913@g.ecc.u-tokyo.ac.jp, ††tomoki_koriyama@ipc.i.u-tokyo.ac.jp

Abstract　In DNN-based speech synthesis, contexts, which are input features of DNN, can be used not only for the representation of linguistic information but also for that of para- and non- linguistic information. Although spontaneous speech synthesis requires the use of various contexts to express the diversity of prosody in spontaneous speech, it is not clear what features are important. In this study, we utilize the rich tags annotated in Corpus of Spontaneous Japanese (CSJ), and use them as the additional contexts. Experimental evaluation results show that both frequently- and infrequently- observed tags are effective for synthesizing spontaneous speech.
Key words　speech synthesis, context, spontaneous speech, annotation, deep neural network

1. ま え が き

音声合成はスマートフォンや対話ロボットにおける音声対話システムだけでなく, 公共交通機関での案内放送や映像作品のボイスオーバーなど, 様々な場面に応用が広がっている. また, この応用場面の拡大に伴い, 与えられたテキストに対応する音声を生成するだけでなく, 多様な感情・発話意図を表現可能な音声合成システムへの期待が高まっている.

統計モデルに基づく音声合成では, 入力されたテキストから音声波形を直接予測することは一般的に困難であるため, テキストから得られるコンテキストと波形生成のための音声特徴量との関係を統計的にモデル化するパイプラインモデルが用いられる. 近年ではテキストの文字列や音素列のみを直接入力とする end-to-end 音声合成を行う研究も広く行われている [1], [2] が, アクセントが重要な日本語 end-to-end 音声合成ではアクセント情報をコンテキストとして加えることで性能が向上するこ

とが報告され [3], コンテキストが依然として重要であるといえる.

本研究では, 統計モデルにおける音声合成の中でも, 深層ニューラルネットワーク (deep neural network: DNN) を用いた音声合成 [4] におけるコンテキストに着目する. 代表的なコンテキストは, 前後の音素情報を示すトライフォンや単語や句に付随する韻律の種類や句の長さなどであり, 読上げ調の DNN 音声合成システムではこのような基本的なコンテキストを用いることで高品質な音声を合成することが可能である. 一方で, コンテキストは自由度が高く様々な情報を付加することが可能であり, 例えば, 文献 [5], [6] では発話者を表すベクトルをコンテキストとして加えることで多話者音声合成システムを実現している. また, 文献 [7], [8] では感情表現の表出度合いをコンテキストとして使用し, DNN 音声合成において感情表現の制御が可能であることを示している. さらに, 強調表現のような局所的な情報もコンテキストによって表現できることが報告されている [9].

以上のことから, コンテキストを加えることによって, より多様な音声を合成できることが期待できる. そこで, 講演や対話などの自発性の高い発話（話し言葉）が収録されている日本語話し言葉コーパス（Corpus of Spontaneous Japanese: CSJ）[10] を用いて, 話し言葉音声合成におけるコンテキストの有効性を検討する. CSJ のうちコアデータと呼ばれる一部の音声データには, 自発音声特有の発話の非流暢性, 音素の引き延ばし, 句末音調の韻律情報などを記述するためのタグが大量に付与されており, 本稿ではこれらのタグを DNN 音声合成のための追加コンテキストとして使用する. このとき, 合成音声の自然音声に対する再現性を主観評価実験により比較し, 話し言葉音声合成に有効なコンテキストを調査する.

2. 関連研究 : HMM 音声合成におけるコンテキストの検討

本研究では DNN に基づく話し言葉音声合成におけるコンテキストの検討を行うが, 隠れマルコフモデル（hidden Markov model: HMM）に基づく話し言葉音声合成 [11] でも同様の研究が行われている [12]. HMM に基づく音声合成ではコンテキストが決定木の分割要素として用いられるため, この研究では, CSJ のタグが決定木分割に与える影響を調査している. 実験結果から, 詳細な韻律を与えるトーンラベルおよび音素の引き延ばしをコンテキストとして用いることで, 自然性が向上することが示されているが, 効果の見られないタグも多く存在した. この理由として, 決定木は上位の分割基準の影響が大きいことや, 条件の排他的論理和が決定木で表現できないことが挙げられる.

本研究で検討する DNN 音声合成では, コンテキスト同士の複雑な関係を自動的にモデル化できる手法であることから, HMM に基づく話し言葉音声合成と比較して, より多様なタグに対して有効性が確認できると考えられる. また, DNN 音声合成では複数話者の同時モデル化が容易に実現可能であるため, 大量の音声による自然性の高い合成音声による評価を行うことができると考えられる.

3. 日本語話し言葉コーパス (CSJ)

日本語話し言葉コーパス (CSJ) [10] は, 現代日本語の自発音声を豊富なアノテーションとともに大量に格納したデータベースである. CSJ の一部であるコアデータには, 137 人の男女による 201 講演が収録されており, 手作業によるものを含め特に集中的にアノテーションが付加されている.

CSJ の XML ファイルは, 図 1 のように詳細なタグにより構成されている [13]. コアデータの XML の階層構造は以下のようになっている.

```
<Talk>
  <IPU>
    <LUW>
      <SUW>
        <TransSUW
          <Mora> or <NonLinguisticSound>
            <Phoneme>
              <Phone>
                <XJToBILabelTone>
                <XJToBILabelWord>
                <XJToBILabelBreak>
                <XJToBILabelPrm>
                <XJToBILabelMisc>
```

XML の要素を用いてコンテキストを形成する方法について述べる.

講演 (Talk): CSJ の音声データは, 講演ごとにファイル化されている. 講演のカテゴリ (学会講演, 模擬講演, 朗読, 対話) や話者の情報はこのタグから得られる.

発話 (IPU): 読み上げ音声においては, 一般に文が発話単位とされる. 一方, 話し言葉音声においては, 必ずしも文末を発声しないので, 文を単位とするのは適切ではない. IPU(inter-pause unit) は, 話し言葉音声における発話単位であり, 200ms 以上のポーズによって区切られる.

長単位 (LUW), 短単位 (SUW): 膠着語である日本語においては, 単語の定義の自由度は高い [14]. それゆえ, CSJ では LUW(long-unit word) と SUW(short-unit word) の 2 種類の単語単位を使用する. 単語の品詞, 活用型, 活用形などの情報は, これらのタグから得られる.

転記短単位 (TransSUW): 転記短単位は, 短単位に非流暢性の情報を付加したものである. 非流暢性の情報と, 一部の単語情報はこのタグから得られる.

モーラ (Mora): このタグには, かな情報が付与されている. 句や節のモーラ数を数えるときや, 句や節の中での位置を数えるときにこのタグの情報が使われる.

NonLinguisticSound: このタグはモーラと同じ階層にあり, 息や笑い声などの非言語的情報のラベルが付与されている.

音素 (Phoneme), 分節音 (Phone): `<Phone>`タグには, 分節音の始端, 終端時刻, 分節音の種類, 無声化母音についての情報が付与されている. 本研究では 59 種の分節音を使用する. 音素は分節音が 1 つ以上連結したもので, 本研究では`<Phoneme>`タグ

図 1　CSJ の XML の例
Fig. 1　Example of XML in CSJ.

は使用しない．

XJToBILABEL*:　これらのタグは，日本語話し言葉の韻律ラベリングスキーム X-JToBI [15] により付与された韻律情報である．<XJToBILabelTone>タグには，句頭境界音調，句末境界音調などの情報が付与されている．<XJToBILabelWord>タグには，語に対するイントネーション情報が付与されており，知覚されたアクセント核の位置情報はこのタグから得られる．<XJToBILabelBreak>タグには，韻律特徴により発話を階層的に構造を表現する BI(break index) 層の情報が付与されている．階層の深度が語境界，アクセント句境界，イントネーション句境界を表すが，このタグに付与されているポーズや非流暢性に関する情報を利用することで，話し言葉音声におけるこれらの境界を柔軟に修正できる．この階層は，4.2 章の XML の階層構造とは別のものである．<XJToBILabelPrm>タグは，<XJToBILabelTone>，<XJToBILabelBreak>タグで解釈できない日本語のイントネーションに対して，補助的にラベル付けされる．<XJToBILabelMisc>タグは，極めて多様な話し言葉音声の韻律に対して，XJToBI で対応できない場合に注釈として付与される．

4.　コンテキスト

4.1　日本語読み上げ音声合成のコンテキスト

読み上げ音声合成のコンテキストの例として，HMM-DNN 音声合成システム (HTS) [16] のデモを取り上げる．日本語のコンテキストでは発話，イントネーション句，アクセント句，モーラ，音素の 5 つの階層的な音声単位を用いる．隣接する単位の影響を考慮するため，現在の単位の情報だけでなく直前および直後の単位の情報をコンテキストに加える．これらのコンテキストに発話者のメタ情報（話者 ID，性別，出生地）と収録講演のカテゴリの情報を加えてベースラインコンテキストとする．

4.2　話し言葉音声合成の追加コンテキスト

話し言葉音声合成の追加コンテキストとして，HMM に基づく音声合成ではトーンラベル，単語単位，節，音素引き延ばし，発話スタイル，非流暢性，音素付加情報が検討された [12]．本稿では，CSJ の XML ファイルに含まれる下記のタグ情報を，追加コンテキストとして使用することを提案する．

4.2.1　トーンラベル

話し言葉音声における音調は読み上げ音声に比べて非常に複雑なので，アクセント情報のみからモデル化するのは困難である．　下降調，上昇調，上昇下降調，下降上昇調，上昇下降上昇調の句末境界音調や，ピッチの上昇・下降の位置を表すポインタは<XJToBILabelTone>タグに付与されている．　これ以外にも，アクセントや句末音調で表現できないピッチ変動を表す<XJToBILabelPrm>タグ，句境界情報を表す<XJToBILabelBreak>タグ，注釈を表す<XJToBILabelMisc>タグの情報を追加コンテキストとして用いる．

4.2.2　単　　語

4.1 章で述べる読み上げ音声合成においては，単語単位の特徴量は重要ではない [17] のでコンテキストから外されている．本研究では，単語単位の特徴量が話し言葉音声において有効かどうかを調査する．品詞，活用型，活用形，音便などの情報は，<LUW>，<SUW>タグから得られる．

4.2.3　節

文末が明示的に現れないことの多い話し言葉音声において IPU 単位は便利であるが，連続的な情報をモデル化するには短すぎることがある．文に関連した文法的単位として，CSJ では転記テキストから自動決定された節情報が<SUW>タグの

ClauseBoundaryLabel に付与されている．本研究では，節情報における境界の強さと境界の種類をコンテキストとして用いる．また, CSJ のコアデータでは人手による節情報の修正が<SUW>タグの CU_OperationSign に付与されているので, コンテキストとして用いる．

4.2.4 重 要 文

話し言葉では要旨となる重要な部分とそうでない部分で発声に差がある． CSJ のコアデータでは，学会講演，模擬講演のうち 177 講演に対して，<SUW>タグの SE_{Subject1|Subject2|Subject3}_{10p|50p}によって, 講演内容に重要な上位 10%, 50% の部分に 3 人の作業者により独立に節単位でラベル付けされている．

4.2.5 音素引き延ばし

話者が考えているとき, 驚いた時, 強調している時, 通常より長く音素を発音することがある．この引き延ばしは字句には表れないので, 別途アノテーションする必要がある．音素引き延ばしの情報は母音, 子音それぞれ<Mora>タグの TagVLong, TagCLong から得られる．

4.2.6 発話スタイル

話者が笑いながら発声したり，囁いたりするとき，音声波形は発話スタイルに依存して変化する．この情報は<Mora>タグ及び<NonLinguisticSound>タグの Tag{Whisper|Laughing|Uncertain|...}から得られる．

4.2.7 非流暢性

話し言葉音声にはさまざまな非流暢な発声が含まれる．語断片, 言い直し, フィラー, 発音エラーの情報が, それぞれ<TransSUW>タグの Tag{Disfluency|Disfluency2|Filler| Incorrect|...}から得られる．

5. 実　　験

5.1 実 験 条 件

CSJ のコアデータに属する 201 講演を使用した．IPU 数は 60424, 音声は 128120 秒（約 35.6 時間）となった．講演を IPU 単位に分割し, CSJ の XML ファイルによって各 IPU にコンテキストラベルを付与した．

DNN は, 一般的なフィードフォワード型ニューラルネットワークとした．隠れ層数は 5 層, 各層のノード数 1024 個で, 活性化関数に ReLU, 勾配法に Adam を使用し, 学習率は 0.001 とした．過学習を防ぐため, 重み減衰係数を 10^{-6}, ドロップアウト率 0.5 とした．ミニバッチサイズは 1024 とし, 20 エポック学習した．

サンプリング周波数 16 kHz の音声波形に対して WORLD [18] による音声分析・合成を行った．音声分析により得られたスペクトル, 非周期性指標, f_o から, 0–59 次のメルケプストラム, 対数 f_o, 1 次元の非周期性指標を抽出した．これらの 1 次および 2 次動的特徴量と有声・無声フラグを合わせた 187 次元のベクトルを音響特徴量とした．入力ベクトルであるコンテキストベクトルは, ベースラインコンテキストのみを用いた場合に, それぞれ音素継続長モデルでは 336 次元, 音響特徴量モデルでは 340 次元とした．

表 1　出現頻度の高いコンテキストについて, ベースラインコンテキストのみの場合とコンテキストを追加した場合を比較した XAB テスト結果
Table 1　XAB test results for additional contexts.

コンテキスト	選択率	p 値
ベースライン vs トーンラベル	45.7 % vs **54.3** %	0.022
ベースライン vs 単語	45.0 % vs **55.0** %	0.014
ベースライン vs 節	44.7 % vs **55.3** %	< 0.01
ベースライン vs 重要文	46.7 % vs 53.3 %	0.10

主観評価用において, 多くの IPU は発話継続長が短いため, IPU を音声サンプルとして用いると聴取者が評価しにくいことが考えられる．そこで本研究では, 主観評価実験の音声サンプルとして, 約 5 秒間の連続する IPU を連結したものを 100 個非復元抽出した．学習用のデータには評価用のデータは含まれないようにした．

5.2 主観評価実験

合成音声によるオリジナル音声の再現性を評価するために, XAB テストに基づく主観評価実験を実施した．クラウドソーシングサービス上で被験者を募集し, 被験者はまずオリジナル音声 X を聞き, 2 種類のコンテキストで合成された A, B のうちどちらが X に近いか選択した．各実験で被験者数は 30 とし, 各被験者は主観評価用データから無作為に 5 つ抽出したものについて, A, B の順序入れ替えを含めて計 10 個評価した．

5.2.1 実験 1：出現頻度の高い追加コンテキストの効果

出現頻度の高いタグから得られるコンテキストについて, ベースラインコンテキストのみの場合と, トーンラベル, 単語, 節および重要文のコンテキストをそれぞれ追加した場合を比較して, その効果を評価した．本実験では, 主観評価用のデータは CSJ のコアデータから無作為に抽出した．表 1 は, ベースラインコンテキストのみの場合とコンテキストを追加した場合を比較した結果である．重要文を除いて, 各追加コンテキストではベースラインコンテキストより有意に高い選択率となっている．

5.2.2 実験 2：コンテキストを複数追加する効果

5.2.1 で有効性が示された追加コンテキストについて, コンテキストを 1 つだけ追加した場合と複数追加した場合を比較して, その効果を評価した．本実験では, 主観評価用のデータは CSJ のコアデータから無作為に抽出した．表 2 は, コンテキストを 1 つだけ追加した場合と複数追加した場合を比較した結果である．コンテキストを 1 つだけ追加した場合よりも, 複数追加した場合の方が高い選択率になっている．特に, 単語または節のコンテキストを追加すると, 有意に高い選択率となっている．

5.2.3 実験 3：出現頻度の低い追加コンテキストの効果

出現頻度の低い音素引き延ばし, 発話スタイル, 非流暢性のタグから得られるコンテキストについて, ベースラインコンテキストのみの場合とコンテキストを追加した場合を比較して, その効果を評価した．本実験で使用する追加コンテキストは, CSJ のコアデータ中でごくわずかしかラベル付けされていない．コンテキストを追加することによる, 追加コンテキストのラベル付けがされている発話への効果とラベル付けがされていない発

表 2　コンテキストを 1 つだけ追加した場合と複数追加した場合を比較した XAB テスト結果
Table 2　XAB test results for the combinations of additional contexts.

コンテキスト	選択率	p 値
トーンラベル vs トーンラベル + 単語	41.0 % vs **59.0** %	$< 10^{-5}$
単語 vs トーンラベル + 単語	46.3 % vs 53.7 %	0.073
トーンラベル vs トーンラベル + 節	41.0 % vs **59.0** %	$< 10^{-5}$
節 vs トーンラベル + 節	43.7 % vs **56.3** %	< 0.01
単語 vs 単語 + 節	41.0 % vs **59.0** %	$< 10^{-5}$
節 vs 単語 + 節	42.0 % vs **58.0** %	$< 10^{-4}$

表 3　出現頻度の低いコンテキストの XAB テスト結果
Table 3　XAB test results for infrequently-observed additional contexts.

コンテキスト	選択率	p 値
ベースライン vs 音素引き延ばし	45.0 % vs **55.0** %	0.014
ベースライン vs 発話スタイル	44.3 % vs **55.7** %	< 0.01
ベースライン vs 非流暢性	43.3 % vs **56.7** %	< 0.01

話への効果を同時に評価するため, 主観評価用データには追加コンテキストのラベル付けがされている発話とラベル付けがされていない発話が同数含まれるようにした. 表3 は, ベースラインコンテキストのみの場合とコンテキストを追加した場合を比較した結果である. 各追加コンテキストではベースラインコンテキストより有意に高い選択率となっている.

6. むすび

本研究では話し言葉音声合成において, 自発音声のように多様な韻律を持つ発話を生成することを目的として, 日本語話し言葉コーパス (CSJ) に付与されているタグを, DNN 音声合成のコンテキストとして用いることの有効性を検討した. 主観評価実験結果より, トーンラベルや節といった出現頻度の高いコンテキストだけでなく, 音素引き延ばしや非流暢性といった出現頻度の低いコンテキストに対しても, 原音声に近い話し言葉音声の合成に有効であることを示した. また, コンテキストを複数追加した場合でも追加コンテキストの組合せは有効であり, さらに, HMM 音声合成における検討 [12] では効果の見られなかった単語や節, 非流暢性といったコンテキストに対しても, DNN 音声合成においては有効でであることを示した.

今後は，さらにコンテキストを加えたときの合成音声の変化や，不要なコンテキストの検討など，詳細な検討を行う予定である.

文　献

[1] J. Sotelo, S. Mehri, K. Kumar, J.F. Santos, K. Kastner, A. Courville, and Y. Bengio, “Char2wav: End-to-end speech synthesis,” 2017.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R.A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” Proc. INTERSPEECH, pp.4006–4010, 2017.

[3] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, “Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language,” Proc. ICASSP, pp.6905–6909, 2019.

[4] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” Proc. ICASSP, pp.7962–7966, 2013.

[5] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, “A study of speaker adaptation for DNN-based speech synthesis,” Proc. INTERSPEECH, pp.879–883, 2015.

[6] N. Hojo, Y. Ijima, and H. Mizuno, “DNN-based speech synthesis using speaker codes,” IEICE Transactions on Information and Systems, vol.E101.D, no.2, pp.462–472, 2018.

[7] S. An, Z. Ling, and L. Dai, “Emotional statistical parametric speech synthesis using LSTM-RNNs,” Proc. APSIPA ASC, pp.1613–1616, 2017.

[8] J. Lorenzo-Trueba, G.E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, “Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis,” Speech Commun., vol.99, pp.135–143, 2018.

[9] M. Wang, Z. Wu, X. Wu, H. Meng, S. Kang, J. Jia, and L. Cai, “Emphatic speech synthesis and control based on characteristic transferring in end-to-end speech synthesis,” 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), pp.1–6, 2018.

[10] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, “Spontaneous speech corpus of japanese.,” Proc. LREC, pp.947–952, 2000.

[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” Proc. EUROSPEECH, pp.2347–2350, 1999.

[12] T. Koriyama, T. Nose, and T. Kobayashi, “On the use of extended context for HMM-based spontaneous conversational speech synthesis,” Proc. INTERSPEECH, pp.2657–2660, 2011.

[13] K. Maekawa, H. Kikuchi, and W. Tsukahara, “Corpus of spontaneous Japanese : Design, annotation, and XML representation,” 2004.

[14] K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka, and Y. Den, “Balanced corpus of contemporary written Japanese,” Lang. Resour. Eval., vol.48, no.2, pp.345–371, 2014.

[15] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. Venditti, “X-JToBI: an extended J-ToBI for spontaneous speech,” Proc. 7th ICSLP, pp.1545–1548, 2002.

[16] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, and K. Tokuda, “The HMM-based speech synthesis system (HTS) version 2.0,” Proc. 6th ISCA Workshop on speech synthesis (SSW6), pp.294–299, 2007.

[17] S. Yokomizo, T. Nose, and T. Kobayashi, “Evaluation of prosodic contextual factors for hmm-based speech synthesis,” Proc. INTERSPEECH, pp.430–433, 2010.

[18] 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, no.7, pp.1877–1884, 2016.

付　録

1. 本稿で使用されているタグと属性

本研究で使用したタグの一覧を表 A·1 および 表 A·2 に記載する. 実験では, 表中のカテゴリごとに, コンテキストを追加した.

表 A·1　タグと属性 (Talk から TransSUW まで).
Table A·1　Tags and attributes (Talk to TransSUW).

タグ・属性	カテゴリ
<Talk>	
TalkID	ベースライン
SpeakerID	ベースライン
SpeakerSex	ベースライン
SpeakerBirthGeneration	ベースライン
SpeakerBirthPlace	ベースライン
<IPU>	ベースライン
IPUStartTime	ベースライン
IPUEndTime	ベースライン
<LUW>	
LUWPOS	単語
LUWConjugateForm	単語
LUWConjugateType	単語
LUWMiscPOSInfo1	単語
LUWMiscPOSInfo2	単語
LUWMiscPOSInfo3	単語
<SUW>	
SUWPOS	単語
SUWConjugateForm	単語
SUWConjugateType	単語
SUWMiscPOSInfo1	単語
SUWMiscPOSInfo2	単語
SUWMiscPOSInfo3	単語
ClauseBoundaryLabel	節
CU__OperationSign	節
SE__Subject1__10p	重要文
SE__Subject2__10p	重要文
SE__Subject3__10p	重要文
SE__Subject1__50p	重要文
SE__Subject2__50p	重要文
SE__Subject3__50p	重要文
<TransSUW>	
TagAlphabetStart	単語
TagAlphabetEnd	単語
TagAlphabetMidst	単語
TagDisfluencyStart	非流暢性
TagDisfluencyEnd	非流暢性
TagDisfluency2Start	非流暢性
TagDisfluency2End	非流暢性
TagFillerStart	非流暢性
TagFillerEnd	非流暢性
TagFillerMidst	非流暢性
TagIncorrectStart	非流暢性
TagIncorrectEnd	非流暢性
TagIncorrectMidst	非流暢性
TagMaskStart	単語
TagMaskEnd	単語
TagMaskMidst	単語
TagForeignStart	単語
TagForeignEnd	単語
TagForeignMidst	単語
TagQuoteStart	単語
TagQuoteEnd	単語
TagQuoteMidst	単語

表 A·2　タグと属性 (Mora から XJToBILabel*まで).
Table A·2　Tags and attributes (Mora to XJToBILabel*).

タグ・属性	カテゴリ
<Mora>	
MoraID	ベースライン
TagVLong	音素引き延ばし
TagCLong	音素引き延ばし
TagWhisperStart	発話スタイル
TagWhisperEnd	発話スタイル
TagWhisperMidst	発話スタイル
TagLaughingStart	発話スタイル
TagLaughingEnd	発話スタイル
TagLaughingMidst	発話スタイル
TagUncertainStart	発話スタイル
TagUncertainEnd	発話スタイル
TagUncertainMidst	発話スタイル
<NonLinguisticSound>	
MoraID	ベースライン
TagBreath	発話スタイル
TagLaugh	発話スタイル
TagVN	発話スタイル
TagWhisperStart	発話スタイル
TagWhisperEnd	発話スタイル
TagWhisperMidst	発話スタイル
TagLaughingStart	発話スタイル
TagLaughingEnd	発話スタイル
TagLaughingMidst	発話スタイル
TagUncertainStart	発話スタイル
TagUncertainEnd	発話スタイル
TagUncertainMidst	発話スタイル
<Phone>	
PhoneEntity	ベースライン
Devoiced	ベースライン
PhoneStartTime	ベースライン
PhoneEndTime	ベースライン
<XJToBILabelTone>	トーンラベル
ToneClass	トーンラベル
Divided	トーンラベル
<XJToBILabelWord>	
PerceivedAccPos	ベースライン
<XJToBILabelBreak>	ベースライン
<XJToBILabelPrm>	トーンラベル
<XJToBILabelMisc>	トーンラベル