Vol.2022-SLP-140 No.31
2022/3/2

個人別自発音声合成の実現に向けた フィラーの言語学的知識に基づく実験的研究

松永 裕太[†1,a)]　佐伯 高明[†1,b)]　高道 慎之介[†1,c)]　猿渡 洋[†1,d)]

概要：本論文では，個人性を再現する自発的な音声合成の実現に向けて，言語学的知識に基づいた包括的な実験的調査を行う．近年発展している音声クローニングは流暢な朗読発話に限定され，より人間らしい自発的な音声合成のための新たな音声クローニングの手法が求められている．そこで本論文は，声色の個人性のみならず非流暢性の個人性を再現可能な自発音声合成に取り組む．具体的には，主要な非流暢性であり，心理学や言語学の研究により発話生成やコミュニケーションにおいて重要な役割を果たすことが知られている，フィラーを扱う．本論文では，話者依存と話者非依存のフィラー予測手法を比較評価するため，多話者コーパスで学習した話者非依存のフィラー予測モデルを用いた音声合成手法を提案する．実験的評価により，フィラーの位置と種類の関連，自然性と個人性のトレードオフを明らかにし，人間らしい音声合成の実現への方向性を示す．

1. はじめに

音声合成は，人間らしい音声を人工的に合成することを目的とする．近年の音声合成は，sequence-to-sequence (seq2seq) モデルの急速な発展により，人間に近い品質の音声を合成可能である [1], [2], [3], [4]．そのような音声合成により話者の個人性を忠実に再現することで，デジタル音声クローニング [5], [6], [7] が実現されている．しかし，これらは流暢な朗読音声の合成を扱っており，自発音声の合成はあまり研究されていない．朗読音声と異なる自発音声の特徴として，繰り返し，言い換え，フィラーなどの非流暢性がある [8]．自発音声合成は，非流暢性を表現することで，テキスト読み上げに留まらないより人間らしい音声を合成する．そのような自発音声合成を個人化できれば，従来の音声クローニングを超える個人別音声合成が可能になる．そこで本研究では，図 1 に示すように，声色の個人性だけでなく非流暢性の個人性も再現する個人別自発音声合成 (personalized spontaneous speech synthesis) の確立を目指す．

フィラーは，言語的内容とは無関係であるが，自発音声において重要な役割を果たす，非流暢性の一種である．フィラーは話者によって発話に挿入されるもので，その役割は言語学的に広く研究されている．例えばフィラーは，話し手が語彙探索中であることを聞き手に伝え [8]，円滑な音声コミュニケーションを可能にする [8], [9], [10]．また，フィラーは聞き手にも影響を与えるとされ，話し手がフィラーを使うことで，聞き手が発話の理解に要する労力（聴取努力）が軽減される傾向が報告されている [11]．このようなフィラーの機能に関する研究だけでなく，話者ごとのフィラーの個人性に関する研究も行われている．例えば，フィラーの単語の選択 [9], [12] や位置 [13] は話者によって変わることが報告されている．自発音声合成において個人性を再現するには，これらの傾向を考慮する必要がある．

自発音声合成を扱った様々な研究 [14], [15], [16], [17] の中でも，AdaSpeech 3 [18] は，フィラーの自動挿入を含む自発音声合成に取り組んでいる．著者らは，音声合成モデルに，入力された流暢なテキストにフィラーを挿入するモデルを導入している．そして，フィラーが挿入されたテキストからフィラーを含む音声を合成している．我々は特に，フィラー挿入方策に着目する．個人別自発音声合成を実現するためには，言語学で解明されたフィラーの傾向（例えば，2.1 節で述べる代替可能性）が，フィラー挿入方策の違いにどのように影響を受けるかを明らかにする必要がある．

本論文では，個人別自発音声合成の実現に向けて，フィラー生成機構を含む自発音声合成モデルを提案し，言語学的知識に基づく実験的評価を行う．我々は，フィラー予測器を備えた Seq2Seq の音声合成モデルを作成する．このモ

[†1] 現在，東京大学大学院情報理工学系研究科
Presently with Graduate School of Information Science and Technology, The University of Tokyo
[a)] matsunaga-yuta339@g.ecc.u-tokyo.ac.jp
[b)] takaaki_saeki@ipc.i.u-tokyo.ac.jp
[c)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
[d)] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

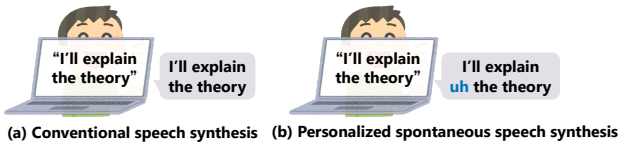

図 1 個人別自発音声合成．声色の個人性のみを再現する従来法（左）に対し，提案法（右）は非流暢性の個人性再現を目指す．

デルは，フィラーのない流暢なテキストから様々な話者に使用される多様なフィラーを予測し，それを挿入したテキストから音声を合成する．実験的評価では，フィラーの言語学的知識に基づいてフィラー予測を評価する．具体的には，フィラーを含む合成音声を，自然性（AdaSpeech 3 [18] と同様）だけでなく個人性や聴取努力を指標として評価する．本論文の貢献は以下の通りである．

- 個人別自発音声合成モデルを提案し，個人別フィラーと非個人別フィラーの比較による包括的な実験的評価を行う．
- 実験的評価により，フィラーの位置，種類の関連と，自然性，個人性のトレードオフを解明し，今後の自発音声合成の研究の方向性を示す．

2. 関連研究

2.1 フィラーに関する言語学的知識

フィラーは，発話プランニング [8], [9] や話者交替 [10] を補助するなど，自発音声において重要な役割を果たす．以下で，個人別音声合成において考慮すべきフィラーに関する主要な言語学的特徴について述べる．

- **語彙：**フィラーの語彙数は膨大である．例えば英語では，“uh” や “um”（AdaSpeech 3 [19] で使用）がよく知られているが，他にも多くの語彙的，非語彙的なフィラー語が存在する [20]．これは他の言語でも同様 [21], [22] で，本論文が対象とする日本語には 160 種類の異なるフィラー語が存在する [23]．
- **代替可能性：**各フィラー語の役割を厳密に分けることはできず，他のフィラー語に変えても効果は変わらないとされる [24]．
- **個人性：**フィラーの様相は，発話内容だけでなく話者によっても変化する [25]．フィラーの位置 [13] と単語 [9], [12] は話者によって異なることが知られている．

2.2 フィラーの合成と評価

自発音声合成の実現を目指し，素片選択 [26] や Seq2seq 型のニューラルネットワーク [15] などを用いた研究がある．しかし，これらの研究はフィラー予測を含まず，フィラー語を含むテキストからの音声合成のみに着目している．[14] は包括的に自発音声合成に取り組み，フィラー予測も行っているが，個人性を考慮していない．フィラー予測を暗黙的 [16] または，明示的 [18] に行う研究では，同様に個人性再現に取り組んでいないだけでなく，フィラー語彙や実験的評価が限定的である．フィラーの位置，種類の予測に特化した研究 [27], [28], [29] では，合成音声の評価は行われていない．これらの先行研究に対し我々は，1) 個人化のための豊富なフィラー語彙を用いた自発音声合成を提案し， 2) 個人別（真）のフィラー挿入と非個人別のフィラー予測による合成音声を言語学的知識に基づいて比較評価する．

フィラーの有無が聞き手による合成音声のパーソナリティ [15] や個人性 [30] の知覚にどのような影響を与えるかを，調査した研究がある．それらは，真（個人別）のフィラーを含む音声と含まない音声を比較している．対して本論文では，話者非依存の予測を用い，フィラーの位置，単語を分けて調査することで，フィラーが個人性知覚に与える影響を詳細に明らかにする．

3. 手法

本節では，新たに設計された語彙を用いた，フィラー予測を含む個人別自発音声合成モデルと，作成した講義音声コーパスについて説明する．我々はこのモデルとコーパスを用いて，4 節の比較評価を行う．

3.1 自発音声合成手法

図 2 に，我々の提案する，話者非依存のフィラー予測モデルと話者依存の音声合成モデルから成る個人別自発音声合成モデルを示す．このモデルは，フィラーのないテキストからフィラーを含む音声を合成する．フィラー予測では，fastText [31] により流暢なテキストの形態素埋め込みベクトル列を取得し，bi-directional long short-time memory (BLSTM) [29] により形態素境界におけるフィラー語を予測する．予測候補は，3.2 節で述べる 13 種類のフィラー語と，“フィラー挿入なし” の計 14 種類である．我々はこのモデルの学習のため，多話者の自発音声から書き起こされた流暢文とフィラーを含む非流暢文から成る，多話者のフィラーのアノテーション付きコーパスを用いる．予測モデルを話者別に学習し当該話者のフィラー語を予測することも考えられるが，本論文では当該話者の真のフィラー語を挿入することでこれを代替する．話者別の予測モデルの学習は今後の課題である．

そして，上記の予測モデルに続き Seq2seq モデルにより音声を合成する．モデルの入力として，音素だけでなく，各音素がフィラーであるかどうかを表す 2 値のフィラータグを用いる．朗読音声による学習済みモデルを，後の 3.3 節で述べるフィラーのアノテーション付き自発音声コーパスを用いて再学習する．

以上の個人別音声合成モデルに，フィラーに関する言語学的知識を導入する．例えば，2.1 節の代替可能性を考慮すると，フィラー語の予測精度は，位置の予測精度に比べ低くても良いことが示唆される．これは，真のフィラー位

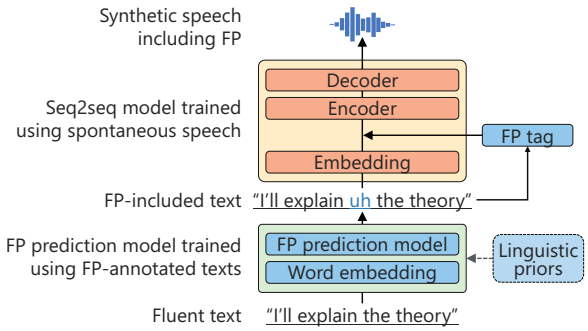

図 2 フィラー予測モデルと音声合成モデルから成る，自発音声合成モデル．4 節で述べるように，言語学的知識をフィラー予測に導入する．

置における真のフィラー語と予測フィラー語の比較により検証される．4 節で詳細を述べる．

3.2 フィラー語彙設計

個人別音声合成のためには，様々な話者の用いる多様なフィラーを含むようフィラー語彙を設計する必要がある．そのため，多話者の自発音声から書き起こされたフィラーのアノテーション付きコーパスを用いて，全話者で使用頻度の低い（< 20%）フィラーを除く．これにより，「えー」「ま」などの 13 種類のフィラー語から成る語彙を作成し，各話者の使用するフィラーをおおよそ 83 %カバーした．

3.3 コーパスとアノテーション

話者依存の音声合成モデルを学習するため，中サイズ（3.5–5.0 時間）の自発音声コーパスを構築する必要がある．本稿では，手動でウェブ上の講義動画を探し，Corpus of Spontaneous Japanese (CSJ [32]) のルールに従ってテキストの書き起こしとアノテーションを行う．コーパスは，書き起こしテキストやフィラーの単語，タグ，時間情報などを含む．著者のホームページで公開されている[*1]．

4. 実験的評価

4.1 実験条件

我々は，フィラー語彙構築とフィラー予測モデル学習のため，CSJ における 137 話者のテキストデータを用いた．語彙構築の詳細は 3.2 節で述べた通りである．

我々は，YouTube 上で講義データを利用可能な 2 名の日本語話者を選定した[*2]．音声データは 16 kHz にダウンサンプリングした．アノテーションは専門家により実施した．以上のデータを，学習，検証，テストデータに分割した．各話者の学習データには，およそ 3.5–5.0 時間のデータが含まれる．テストデータには，20 個の音声セグメントを用いた．各音声セグメントは複数文から成り，およそ 15–30 秒の長さである．

フィラー予測モデルは，Common Crawl[*3]と Wikipedia で学習済みの fastText [31] と，BLSTM から成る．学習データは，25 時間の自発音声から書き起こされたおよそ 35,000 の呼気段落である．形態素解析には Sudachi [33] を用いた．BLSTM の隠れ層の数，次元数は，それぞれ 1, 1024 とした．ノルムの最大値を 0.5 とする勾配クリッピングを適用し，バッチサイズを 32 とした．学習率を 1.0×10^{-3} とし， 20 エポックごとに 0.1 倍した．これらの条件でモデルを 400,000 ステップ学習させた．

音声合成モデルは FastSpeech 2 [2] の構造を基にした．まず，JSUT コーパス [34]（女性日本語話者による 10 時間の読み上げ音声）を用いて事前学習し，アノテーションされた自発音声を用いて再学習した．モデル構造や事前学習の条件は公開実装に従った[*4]．自発音声を呼気段落に分割し，500,000 ステップ再学習させた．音素ごとの 2 値のフィラータグを音素埋め込みに結合し，エンコーダの入力とした．合成において，音声を呼気段落ごとに独立に合成し，結合することにより音声セグメントとした．

4.2 評価

合成音声におけるフィラー予測の効果を調べた．2.1 節で述べた代替可能性を考慮するため，フィラーを位置と語の要素に分割して調査した．表 1 に手法の一覧を示す．学習されたモデルは手法間で同一であり，音声合成モデルの入力のみが異なる．

NoFP は，入力テキストにフィラーを含まず，流暢な音声を合成する．**PredFP** は，CSJ コーパスにおける各種フィラーの生起確率を基にフィラーをランダムに挿入する．**PredFP-TruePos** は，真のフィラー位置を参照し，その位置におけるフィラー語のみを予測する．**TrueFP** は，真のフィラー位置，語の両方を参照する（すなわち，TrueFP のフィラー挿入テキストは，真のテキストと完全一致する）．

以上で述べた手法を用い，合成音声に対する主観評価を行った．我々は個人別音声合成の実現を目的としているため，自然性 (naturalness) だけでなく個人性 (individuality) を評価指標として用いた．また，フィラーの重要な役割 [11] であり，「聴取するために必要な努力がどのくらい少ないか？」を表す評価指標である，聴取努力 (listening effort) についても評価した．

以下では特に述べない限り，評価形式を AB，XAB テストとする．個人性の XAB テストにおける “X” は，ground-truth の音声（話者の自然音声）を表す．各テストには，合計 30 人の聴取者が参加した．各聴取者は，AB テストでは 10 組，XAB テストでは 8 組の音声サンプルを評価し

[*1] https://sites.google.com/g.ecc.u-tokyo.ac.jp/yuta-matsunaga/publications/spon_utokyo_lecture
[*2] https://youtube.com/playlist?list=PLHxBhbJJasnfX6oBrkTygP8we61wEcRha
[*3] https://commoncrawl.org/
[*4] https://github.com/Wataru-Nakata/FastSpeech2-JSUT/tree/master/config/JSUT, commit 549edd9 (model-jdit), fa0aed5 (model, train), 8131cff (preprocess).

表 1　評価において比較する手法の一覧

Name	FP word	Position	Example
NoFP	–	–	I'll explain the theory.
RandFP	Random	Random	I'll explain the theory uh.
PredFP	Predicted	Predicted	I'll uh explain the theory.
PredFP-TruePos	Predicted	Ground-truth	I'll explain um the theory.
TrueFP	Ground-truth	Ground-truth	I'll explain uh the theory.

表 2　NoFP と TrueFP を比較した主観評価結果

Criterion	Spk.	Score NoFP vs. TrueFP	p-value
Naturalness	A	**0.573** vs. 0.427	3.13×10^{-4}
	B	**0.627** vs. 0.373	3.07×10^{-10}
Individuality	A	**0.550** vs. 0.450	2.85×10^{-2}
	B	**0.579** vs. 0.421	4.98×10^{-4}
Listening effort	A	**0.563** vs. 0.437	1.88×10^{-3}
	B	**0.580** vs. 0.420	8.27×10^{-5}

表 3　PredFP と RandFP を比較した主観評価結果

Criterion	Spk.	Score PredFP vs. RandFP	p-value
Naturalness	A	**0.837** vs. 0.163	$<10^{-10}$
	B	**0.840** vs. 0.160	$<10^{-10}$
Individuality	A	**0.758** vs. 0.242	$<10^{-10}$
	B	**0.863** vs. 0.137	$<10^{-10}$
Listening effort	A	**0.770** vs. 0.230	$<10^{-10}$
	B	**0.783** vs. 0.217	$<10^{-10}$

た．音声サンプルは，20 の音声セグメントから成るテストセットからランダムに選択した．

4.2.1　フィラー合成の評価

我々はまず，合成音声における自然性，個人性，聴取努力が，合成されたフィラーにどのように影響を受けるかを調査した．ここで，フィラー予測モデルの性能による影響を避けるため，真のフィラー位置と語を用いる．表 2 に，“NoFP” と “TrueFP” を比較した主観評価の結果を示す．太字は，$p < 0.05$ で有意に優れるスコアである．この表記は，以降の AB テスト，XAB テストの結果の表においても同様である．全ての指標について，フィラーを含む音声よりもフィラーのない音声の方が良い結果になっている．これより，音声合成モデルがフィラーを考慮しても，フィラーを考慮しない場合と比べて結果が改善しないことが分かる．

4.2.2　フィラー予測の評価

提案法におけるフィラー予測の効果を調べるため，主観評価により “PredFP” と “RandFP” を比較した．表 3 に結果を示す．全ての指標について，ランダムなフィラーが挿入された場合と比べ，フィラー予測によって合成音声の評価が有意に改善した．これより，自発音声合成においてフィラー予測が効果的であることが分かる．

4.2.3　フィラーの位置と単語予測の評価

フィラーの位置と語を予測した合成音声と，真のフィラー位置および語を使用した合成音声を比較評価した．表 4 に，“PredFP” と “TrueFP” を比較した主観評価の結果を示す．TrueFP は，話者 A の個人性評価において有意に高く評価されたが，PredFP は全ての評価で自然性，聴取努力について有意に高く評価された．これらの結果は，フィラーを予測することで話者の個性とは関係なく自然性と聴取努力を向上させることができる一方，個人性のような話者依存の指標については，話者依存のフィラーを用いた方が良いことを示している．

表 4　PredFP と TrueFP を比較した主観評価結果

Criterion	Spk.	Score PredFP vs. TrueFP	p-value
Naturalness	A	**0.557** vs. 0.443	5.45×10^{-3}
	B	**0.553** vs. 0.447	8.93×10^{-3}
Individuality	A	0.442 vs. **0.558**	1.05×10^{-2}
	B	0.500 vs. 0.500	1.00
Listening effort	A	**0.553** vs. 0.447	8.93×10^{-3}
	B	**0.543** vs. 0.457	3.38×10^{-2}

表 5　PredFP-TruePos と TrueFP を比較した主観評価結果

Criterion	Spk.	Score PredFP-TruePos vs. TrueFP	p-value
Naturalness	A	0.460 vs. 0.540	5.02×10^{-2}
	B	0.450 vs. **0.550**	1.43×10^{-2}
Individuality	A	0.421 vs. **0.579**	4.98×10^{-4}
	B	0.462 vs. 0.538	1.01×10^{-1}
Listening effort	A	0.480 vs. 0.520	3.28×10^{-1}
	B	0.420 vs. **0.580**	8.27×10^{-5}

4.2.4　フィラーの単語予測の評価

フィラー語をより正確に予測することの重要性を調べるため，“PredFP-TruePos” と “TrueFP” を比較した．表 5 に結果を示す．表を見ると，全ての評価指標において，片方の話者で真のフィラー位置および語を使用した方が有意に高い評価となり，もう一方の話者で同等の評価となっている．これは，フィラーの位置が既知の場合は，フィラー語を予測するよりも真のフィラー語を用いる方が各指標が向上することを示唆している．よって，フィラー語の予測精度を向上させることで，合成音声の評価を向上させることができることが示唆された．

4.2.5　フィラーの位置予測の評価

フィラーの位置が合成音声の評価に与える影響を調べるため，“PredFP” と “PredFP-TruePos” を比較した．表 6

表 6 PredFP と PredFP-TruePos を比較した主観評価結果

Criterion	Spk.	Score PredFP vs. PredFP-TruePos	p-value
Naturalness	A	**0.590** vs. 0.410	9.16×10^{-6}
	B	**0.603** vs. 0.397	3.27×10^{-7}
Individuality	A	**0.604** vs. 0.396	4.16×10^{-6}
	B	**0.600** vs. 0.400	1.01×10^{-5}
Listening effort	A	**0.603** vs. 0.397	3.27×10^{-7}
	B	**0.570** vs. 0.430	5.84×10^{-4}

表 7 NoFP と PredFP を比較した主観評価結果

Criterion	Spk.	Score NoFP vs. PredFP	p-value
Naturalness	A	0.490 vs. 0.510	6.25×10^{-1}
	B	0.493 vs. 0.507	7.44×10^{-1}
Individuality	A	**0.596** vs. 0.404	2.36×10^{-5}
	B	**0.550** vs. 0.450	2.85×10^{-2}
Listening effort	A	**0.583** vs. 0.417	4.08×10^{-5}
	B	**0.507** vs. 0.493	7.44×10^{-1}

に結果を示す．表を見ると，全てにおいて予測されたフィラーを使用した方が有意に高い評価となっており，真のフィラー位置を用いるだけでは，個人性を再現するのに不十分であることを示している．

4.2.6 フィラーを含む最良の手法の評価

これまでの評価において合成されたフィラーを用いている手法の中で最も良い “PredFP” と，フィラーを含まない手法である “NoFP” を比較した．表 7 に結果を示す．自然性を除く全てにおいて，フィラーを含まない手法が有意に高い評価となっている．また自然性については，フィラーを予測し合成した手法がフィラーなしの手法と同等の結果となっている．

4.2.7 絶対評価

最後に，絶対評価実験を実施した．他の手法より評価の低い “PredFP-TruePos” を除外し，比較のため ground-truth となる自然音声を追加した．自然性，聴取努力に対し 5 段階の Mean Opinion Score（MOS）テストを，個人性に対し ground-truth の音声を参照した 5 段階 Degradation MOS（DMOS）テストを実施した．各試験には合計 100 名の聴取者が参加した．各聴取者は，MOS テストで 15 個，DMOS テストで 12 個の音声サンプルを評価した．

表 8 に結果を示す．この結果は，これまでのプリファレンステストの結果と一致している．流暢な音声（NoFP）のスコアは 3.0–3.3 程度であり，自然性のスコアは現在の読み上げ音声よりもまだ低い．また，非流暢な音声のスコア（PredFP や TrueFP など）は流暢な音声と同等なものもあるため，自発音声合成の基本性能の改善により本手法全体の性能が向上することが示唆された．

表 8 合成音声の Mean opinion score (MOS)．“conf” は 95% 信頼区間．

Criterion	Method	Mean ± conf
Naturalness	NoFP	3.291 ± 0.094
	TrueFP	3.028 ± 0.095
	PredFP	3.274 ± 0.096
	RandFP	1.865 ± 0.090
	Ground-truth	4.294 ± 0.091
Individual	NoFP	3.170 ± 0.109
	TrueFP	3.192 ± 0.109
	PredFP	3.115 ± 0.112
	RandFP	2.061 ± 0.111
Listening effort	NoFP	3.043 ± 0.101
	TrueFP	2.884 ± 0.097
	PredFP	3.035 ± 0.100
	RandFP	1.800 ± 0.101
	Ground-truth	3.974 ± 0.091

4.3 考察

本節では，4.2 節で行った評価実験について考察する．まず，4.2.1 節で述べたように，位置，語ともに真のフィラーを用いた場合，フィラーを用いない場合よりも合成音声の品質が低く，これは先行研究 [15] の結果と一致する．さらに，フィラーの位置と語を予測し合成する方法は，真の位置と語のフィラーを使用する方法よりも自然性が優れていた．フィラーを予測する方法は，CSJ コーパスで学習した個人性を考慮しない予測モデルを用いている．そのような話者非依存のフィラー予測は，個人性を考慮したフィラー予測よりも合成音声の自然性を向上させ，フィラーを用いない手法と同等の自然性を実現することができる．一方，個人性については，真の位置と語のフィラーを用いた場合の話者 A における評価が有意に高かった．このことから，今後，個人性を正確に再現したフィラー予測モデルを学習することで，高い個人性再現度のフィラー合成が可能になることが示唆される．

5. 結論

我々は，フィラー予測を含む個人別自発音声合成を提案した．言語学的知識に基づきフィラー予測の効果を調べ，今後の個人性を考慮した自発音声合成の実現のための方向性を示した．今後は，自発音声合成の基本性能の向上と自発音声コーパスの自動構築に取り組む必要がある．

謝辞 本研究は，JST ムーンショット型研究開発事業 JPMJMS2011 の支援を受けたものです．

参考文献

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, Calgary, Canada, 4 2018, pp. 4779–4783.

[2] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2021.

[3] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *Proc. ICASSP*, Online, 6 2021, pp. 5679–5683.

[4] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," in *Proc. ICLR*, 2021.

[5] Q. Xie, X. Tian, G. Liu, K. Song, L. Xie, Z. Wu, H. Li, S. Shi, H. Li, F. Hong, H. Bu, and X. Xu, "The multi-speaker multi-style voice cloning challenge," in *Proc. ICASSP*, 2021, pp. 8613–8617.

[6] M. Blaauw, J. Bonada, and R. Daido, "Data efficient voice cloning for neural singing synthesis," in *Proc. ICASSP*, 2019, pp. 6840–6844.

[7] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, 2018.

[8] W. J. Levelt, "Monitoring and self-repair in speech," *Cognition*, vol. 14, no. 1, pp. 41–104, 1983.

[9] S. Elisabeth, "Preliminaries to a theory of speech disfluencies," *Unpublished PhD dissertation, University of California, Berkeley*, 1994.

[10] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[11] J. E. Arnold, M. K. Tanenhaus, R. J. Altmann, and M. Fagnano, "The old and thee, uh, new: Disfluency and reference resolution," *Psychological Science*, vol. 15, no. 9, pp. 578–582, 2004.

[12] M. Watanabe and Y. Shirahata, "Factors related to probabilities of clause-internal "ee", "anoo" and "maa" in simulated public speaking of csj," in *Proc. Language Resources Workshop*, vol. 4, 2019, pp. 359–367, in Japanese.

[13] E. Shriberg, "Disfluencies in switchboard," in *Proc. ICSLP*, vol. 96, no. 1. Citeseer, 1996, pp. 11–14.

[14] R. Dall, "Statistical parametric speech synthesis using conversational data and phenomena," *PhD dissertation of the University of Edinburgh*, 2017.

[15] J. Gustafson, J. Beskow, and E. Szekely, "Personality in the mix - investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis," in *Proc. 11th ISCA SSW*, 2021, pp. 48–53.

[16] Éva Székely, G. Eje Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," in *Proc. 10th ISCA SSW*, 2019, pp. 245–250.

[17] Y. Yamashita, T. Koriyama, Y. Saito, S. Takamichi, Y. Ijima, R. Masumura, and H. Saruwatari, "Investigating Effective Additional Contextual Factors in DNN-Based Spontaneous Speech Synthesis," in *Proc. Interspeech*, 2020, pp. 3201–3205.

[18] Y. Yan, X. Tan, B. Li, G. Zhang, T. Qin, S. Zhao, Y. Shen, W.-Q. Zhang, and T.-Y. Liu, "Adaptive Text to Speech for Spontaneous Style," in *Proc. Interspeech*, 2021, pp. 4668–4672.

[19] Y. Yan, X. Tan, B. Li, T. Qin, S. Zhao, Y. Shen, and T.-Y. Liu, "Adaspeech 2: Adaptive text to speech with untranscribed data," in *Proc. ICASSP*, 2021, pp. 6613–6617.

[20] G. Brown, *Listening to Spoken English*, ser. Applied Linguistics and Language Study. Taylor & Francis, 2017.

[21] S. Strassel, J. Kolár, Z. Song, L. Barclay, and M. Glenn, "Structural metadata annotation: moving beyond English," in *Proc. Interspeech*, 2005, pp. 1545–1548.

[22] Y. Zhao and D. Jurafsky, "A preliminary study of mandarin filled pauses," in *Disfluency in Spontaneous Speech*, 2005.

[23] K. Hirose, Y. Abe, and N. Minematsu, "Detection of fillers using prosodic features in spontaneous speech recognition of Japanese," in *Proc. Speech Prosody*, 2006, p. paper 187.

[24] K. Yamashita and E. Mizukami, "Using fillers as mental makers: Effects of familiarity, modality, and task difficulty in describing the figure," *Journal of Natural Language Processing*, vol. 14, no. 3, pp. 39–60, 2007, in Japanese.

[25] H. H. Clark and T. Wasow, "Repeating words in spontaneous speech," *Cognitive Psychology*, vol. 37, no. 3, pp. 201–242, 1998.

[26] J. Adell, A. Bonafonte, and D. Escudero-Mancebo, "On the generation of synthetic disfluent speech: local prosodic modifications caused by the insertion of editing terms," in *Proc. Interspeech*, 2008, pp. 2278–2281.

[27] K. Ohta, M. Tsuchiya, and S. Nakagawa, "Construction of spoken language model including fillers using filler prediction model," in *Proc. Interspeech*, 2007, pp. 1489–1492.

[28] M. Tomalin, M. Wester, R. Dall, B. Byrne, and S. King, "A lattice-based approach to automatic filled pause insertion," in *DiSS The 7th Workshop on Disfluency in Spontaneous Speech*, 2015.

[29] Y. Yamazaki, Y. Chiba, T. Nose, and A. Ito, "Filler prediction based on bidirectional lstm for generation of natural response of spoken dialog," in *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, 2020, pp. 360–361.

[30] Éva Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Spontaneous Conversational Speech Synthesis from Found Data," in *Proc. Interspeech*, 2019, pp. 4435–4439.

[31] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[32] K. Maekawa, "Corpus of spontaneous japanese : its design and evaluation," *Proc. SSPR*, pp. 7–12, 2003.

[33] K. Takaoka, S. Hisamoto, N. Kawahara, M. Sakamoto, Y. Uchida, and Y. Matsumoto, "Sudachi: a japanese tokenizer for business," in *Proc. LREC*, may 2018, pp. 2246 – 2249.

[34] R. Sonobe, S. Takamichi, and H. Saruwatari, "Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," *ArXiv*, vol. abs/1711.00354, 2017.