意味の理解に不要な挿入語を検出して除去する話し言葉音声認識エンジン

Speech Recognition Engine with Detection and Removal of Filler Words Unnecessary for Understanding Spontaneous Speech

藤村　浩司　■FUJIMURA Hiroshi
益子　貴史　■MASUKO Takashi
永尾　学　■NAGAO Manabu

近年，深層学習（ディープラーニング）技術の導入により，音声の認識性能は飛躍的に向上している。しかし，人が日常のコミュニケーションのために話す話し言葉音声の認識性能は，人が機械などを意識した読上げ音声に比べて，大幅に劣化する。認識性能の劣化要因は複数存在するが，会話中に発声される「えーと」，「あの」，「まー」といった場つなぎのフィラーと呼ばれる挿入語による性能の劣化は，音声認識システムで解決されていない課題の一つである。またフィラーが正しく認識されたとしても，音声認識結果から意味を理解するうえで妨げになるという問題も存在する。

東芝は，話し言葉音声の認識性能を向上させ，ユーザーが意味を理解しやすい結果を出力するために，高精度な話し言葉音声認識エンジンの開発に取り組んでいる。今回，音響イベント検出と音韻識別を同時に行う音響モデルを利用して，フィラーを検出して除去できる音声認識のデコーディングアルゴリズムを開発し，当社の音声認識エンジンに実装した。評価実験の結果，この音声認識エンジンがフィラーを高精度に検出して除去できるとともに，話し言葉に対して高精度に動作することを確認した。

The performance of speech recognition has recently become increasingly sophisticated due to the application of deep learning technologies. However, the performance level in the case of recognition of spontaneous speech in daily communications is much lower than that of read speech including human-to-machine speech. Performance deterioration associated with filler words inserted into spontaneous speech, such as *um* and *ah,* is a serious technical issue that still remains to be solved. Furthermore, these filler words hinder understanding of the meaning of speech recognition results.

Toshiba has been actively focusing on the development of a speech recognition engine with enhanced robustness to spontaneous speech in order to provide users with easy-to-understand speech recognition results. We have now developed a decoding algorithm for speech recognition capable of detecting and removing filler words by applying an acoustic model that can simultaneously implement acoustic event detection and phoneme recognition, and have incorporated this algorithm into our speech recognition engine. We have conducted verification tests and confirmed the robustness of this speech recognition engine to spontaneous speech as well as high-performance filler word detection and removal.

1　まえがき

東芝は，1960年代から音声や，映像及び画像，言語，知識などの処理技術を研究開発しており，現在は“東芝コミュニケーションAI RECAIUS™”の中で，様々な音声認識ソリューションを展開している。その中には，会議での発言やコールセンターでの会話を認識し，長時間の会議内容を把握したり，オペレーター応答業務を支援したりするといった，話し言葉音声認識技術の応用がある。音声認識技術は近年，深層学習技術の導入により，使用場面や使用環境を限定すれば実用的な音声認識レベルに達するものも出てきた。しかし，例えば議論中心の会議などのように，話し言葉音声を認識する応用では，まだまだ課題がある。話し言葉音声の認識では，「えーと」，「あの」，「まー」といった場つなぎの挿入語であるフィラーや，言いよどみ，言い誤り，発声の怠け，早口などの原因により，音声認識の精度が著しく低下する[1]。また，音声認識結果にフィラーや，言いよどみ，言い誤りなどの語の断片が挿入されると，音声認識結果の可読性も低下する。

当社は，話し言葉音声の認識精度を向上させるために，特に影響度が大きいフィラーへの対策に着目し，これまでにフィラーラベルを付与した単語辞書を用いて，フィラー検出が可能な音響モデルを構築した[2]。今回，その音響モデルを用いて，より自由度の高いフィラー検出が可能なデコーディングアルゴリズムを開発し，それを当社の音声認識エンジンに実装した。またこの音声認識エンジンを用いて，フィラー検出精度と話し言葉音声の認識精度の評価を行った。

ここでは，今回開発した高精度なフィラー検出方式の概要と，その方式を用いたフィラー検出精度及び話し言葉の音声認識精度の評価結果について述べる。

2　音声認識エンジン

通常，音声認識エンジンは，大きく三つの部分に分けられる。一つ目は音響モデルと呼ばれ，音声波形をフレーム単位で区切って，そのフレームがどのような音素と音節であるかを識別する部分である。二つ目は言語モデルと呼ばれ，単語の連鎖確率をモデル化した部分である。例えば，「明日の天気は」の次に来る単語の確率を，（晴れ，0.3），（曇り，0.3），（雨，0.3），

（良い，0.05），（悪い，0.05）などのように表す。三つ目はデコーダーと呼ばれ，音響モデルと言語モデルのスコアを統合し，膨大な単語の組合せの中から音声入力に対して最適な単語列を選択する部分である。

今回開発した方式では，音響モデルとデコーダーの工夫によってフィラーの検出・除去性能を向上させた。

3　フィラー対応音声認識

フィラーは他の単語として誤って認識されると，音声認識精度が大きく低下する。音声認識結果の可読性を向上させるためには，認識結果からフィラーを除去することが望ましい。今回開発した方式の音声認識エンジンの概要を，**図1**に示す。

3.1　フィラー検出音響モデル

これまでに当社が提案した，音響モデル[2]で単語辞書に登録したフィラーを検出する手法では，音響モデルの学習に用いるラベル（音声に対して単語や音素を対応付けしたもの）としてフィラーラベルを導入し，各フィラーの末尾にそれを付加する。例えば，フィラーラベルをFとすると，「（ええっとF），今日の，（あのF），課題は」のようにラベルを付ける。ここで，ひらがな1文字に相当する音節を識別するようにモデルを構築すると，音響モデルの学習に使用するラベルは「ええっとFきょうのあのFかだいわ」となる。

音声に対して，正解の音節列それぞれが音声のどの部分に当たるかのアライメントをあらかじめ取り，フィードフォワード型のニューラルネットワークをそのアライメントを基に学習する方法がある。この方法では，音声波形を一定区間切り出したときの1フレームにつき，必ず一つの音節ラベルが割り当てられる。しかし，Fのラベル部分に相当するアライメント部分は時間的に独立に存在しない。例えば，「あのF」ではフィラー自体は「あの」であるが，この時間区間は「あ」と「の」のラベルが割り当てられ，「F」を割り当てることができない。したがって，フレームとラベルを1対1に対応させる方法では学習が困難であった。

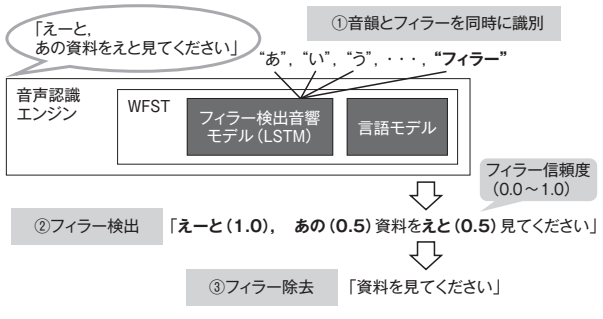


図1．フィラーを検出して除去する音声認識エンジンの概要 ― 今回開発した方式では，フィラーを高精度に検出して除去できる。
Outline of speech recognition engine with detection and removal of filler words

この問題を解決するために，当社は，長時間の依存関係をモデル化できる深層学習技術の一種であるLSTM（Long Short-Term Memory）と，時間軸上でラベル位置を推定しながら学習を進めるCTC（Connectionist Temporal Classification）[3]を導入した。これらの特性を生かして，従来の音声認識を行いながら，フィラーラベルを検出する手法を確立した。

しかし，音響モデル側でこれらのラベルを検出できたとしても，従来の単語辞書を用いてデコードする場合には，単語辞書側にフィラーラベルFを付随したフィラーを用意しておく必要がある。例えば，「ええっとF」，「あのF」などを全て単語辞書に網羅しておかなければならない。また音響的にはフィラーを検出しても，前後の単語との関係によっては言語モデルのスコアに従って他の単語に変換されてしまうことがある。そこで，この問題をデコーダー側の工夫によって解決するとともに，フィラーかどうかの確からしさを表すフィラー信頼度も付与することにした。

3.2　音声認識デコーダー

当社は，図1のフィラー検出音響モデルと言語モデルを用いてスコアによる最適系列の探索を効率良く行うことができる，重み付き有限状態トランスデューサー（WFST）をベースに，当社オリジナルのデコーダーを開発した。これには当社独自の方式が多く組み込まれており，その中の主な特長について述べる。

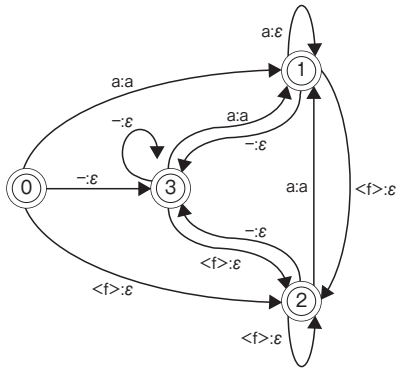
WFSTを用いた従来のデコード方式では，ノードと，ノード間を結ぶことで経路を構成するアークを生成及び破棄しながら探索を行うが，当社方式では，アークは生成せずにノードから必要に応じてアークを復元する。これにより，アークの生成及び破棄に掛かる計算量とメモリー使用量を削減できる[4]。

このデコーダーをベースとして，単語辞書にフィラーラベルFを付与したフィラーを追加することなく，更にフィラー信頼度を付与したフィラー検出アルゴリズムを開発し，音声認識デコーダーに実装した。

3.3　フィラー検出音声認識デコーダー

フィラーラベルを発音の一部として単語辞書に組み込んでいた従来手法とは異なり，今回開発したフィラー検出音声認識デコーダーでは，フィラーラベルを単語の発音記号としては利用しない。このアルゴリズムは，単語辞書に存在する単語を構成する音節ネットワークの経路探索中に，音響モデルが出力するフィラーラベルを受け付けた場合にその単語をフィラーとみなす，といった動作をする。具体的には，**図2**に示すWFSTを構築することで実現する。図2で表されるネットワークは，単語を構成する音節記号aが入力される場合は基本的に変換a:aによってそのまま出力し，単語中へのフィラー入力はフィラー記号<f>から空記号εに<f>:εによって変換する。これによって出力される単語系列はそのままで，<f>の有無に

一般論文



a：音節記号の一つ　　−：ブランク

＊1：⓪, ①, ②, 及び③は，状態を表す
＊2：a:aは音節記号を同じ音節記号へ変換することを，
a:εは音節記号を空記号へ変換することを，
−:εはブランクを空記号へ変換することを，
<f>:εはフィラー記号を空記号へ変換することを，それぞれ表す

図2．フィラー処理WFSTの例 ― フィラー検出デコーディングのために，従来のWFSTにフィラー処理用WFSTを追加で合成する。

Example of weighted finite-state transducer (WFST) to detect filler words

よってその単語がフィラーかどうかを判断できるようになる。

デコーダーは，WFSTの経路上の入力記号を見ることで各単語がフィラーであるか否かを判断する。従来の音声認識においては，探索したWFSTの経路上にある出力記号をつなげて認識結果としている。このときデコーダーは，その経路上の入力記号も参照できる。経路上の入力記号には，図2のWFSTの入力記号が現れるので，単語wに対応する経路上の入力記号列に<f>が含まれるとき，単語wはフィラーであるとして認識結果を出力する。

このようなWFSTを従来の音声認識に使用するWFSTと合成することで，単語辞書にフィラーラベルが付与されたフィラーを語彙として網羅していなくても，フィラーを検出できる。

3.4　フィラー信頼度

フィラー検出音声認識デコーダーとフィラー検出音響モデルを用いて，フィラー信頼度を算出する方法について述べる。

これまでに当社が提案したフィラー検出音響モデル[2]では，「ええっとF」のように単語末にフィラーラベルのFを挿入して学習していたが，フィラー信頼度を導入するために，「えFえFっFとF」のようにフィラーの音節ごとにフィラーラベルを挿入して学習する。このように学習すると，単語wに対応する経路上の入力記号列に複数の<f>が現れる。その語がフィラーであれば，学習時と同じように音節の数と同じ<f>が現れるはずである。これを利用して，その単語の音節数とデコード中に現れる入力記号列<f>の数によってフィラー信頼度を算出する。例えば，「ええっと」のように4音節のフィラーに対して，<f>が二つ検出された場合にフィラー信頼度は50％となり，<f>が四つ検出された場合には100％となる。

この手法を用いることで，頻出するフィラーを部分系列に持つ長い単語を，誤ってフィラーと判断してしまうのを防ぐことができる。例えば，「グラデュエート」などに含まれる「エート」などに誤って<f>が検出されても，「グラデュ」部分には<f>が発生しにくいためフィラー信頼度は低くなる。

4　評価実験

4.1　フィラー検出精度の評価尺度

今回開発した方式を用いたフィラー検出精度と，話し言葉の音声認識精度の評価実験について述べる。フィラー検出の正誤判定は，正解フィラー区間とフィラーであると認識された検出フィラー区間について論理和（OR）区間に対する論理積（AND）区間の割合に基づいて行う。AND区間は正解フィラー区間かつ検出フィラー区間である区間，OR区間は正解フィラー区間又は検出フィラー区間である区間である（**図3**）。

AND区間の区間長をc，OR区間の区間長をlとしたとき，パラメーター$(l\text{-}c)/l$が許容値以下の場合に正しく検出されたと判定する。許容値が大きいほど，AND区間の割合が小さくても正解と判定される。この判定に基づいて，正解フィラー区間を正しく検出できなかった割合の誤棄却率FRと，1時間当たりの誤検出回数FAが求められる。

このとき，LSTMはCTCに基づいて学習されているため，アライメントに基づいて学習されるDNN（Deep Neural Network）とは異なり，出力記号が出力される時間が音声波形上の対応する音響的なイベントの時間と一致しないという問題がある。実際に音声波形と出力記号の出力確率の系列を比較すると，出力記号が音響的なイベントから遅れて出力される傾向が見られることがわかる。そこでフィラー検出精度のパラメーターとして，時間的なずれを補正する時間オフセットを追加し，検出フィラー区間を時間オフセット分だけずらしてから正誤判定を行うことにした。

4.2　フィラー検出精度の評価

講演を接話マイクで収録した日本語話し言葉コーパス（CSJ：Corpus of Spontaneous Japanese）の認識評価セットであるCSJテストセット3[5]を用いて，フィラー検出精度を評価した。

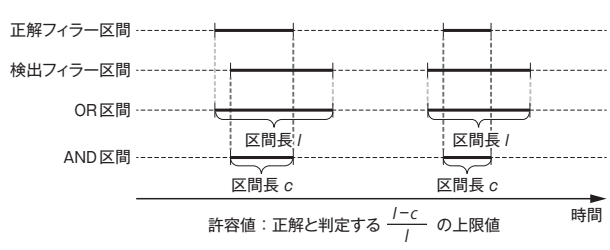


図3．許容値の定義 ― フィラー検出精度の評価尺度として，正解フィラー区間と検出フィラー区間のずれの許容値を定義した。

Definition of tolerance for filler word detection section

フィラー検出精度の評価パラメーターには，フィラーの信頼度，許容値，及び時間オフセットの三つがある。

まず，許容値を0.5として時間オフセットとフィラー検出精度との関係を検討した。その結果を**図4**に示す。図中の数字はフィラー信頼度に対するしきい値であり，しきい値以上の区間だけを検出フィラー区間として扱うことを表している。フィラー信頼度のしきい値に関わらず時間オフセットが-0.34 s，すなわち検出フィラー区間を0.34 s前にずらした場合に*FR*がもっとも低くなっていることがわかる。そこで，以下では時間オフセットを-0.34 sとして評価を行うことにした。

次に，時間オフセットを-0.34 sとして，許容値とフィラー検出精度との関係を調べた結果を，**図5**に示す。許容値が大きいほど正解とみなされる区間が増加するため，*FR*は減少するが0.8～0.9程度で収束していることがわかる。

フィラー信頼度のしきい値を0から100まで10刻みで変化させた場合の*FR*と1時間当たりの*FA*を許容値ごとにプロットしたものを**図6**に示す。図中の各曲線の右端がフィラー信頼度のしきい値0，左端が100の点を表す。フィラーラベルが付与されている単語及びそのフィラー信頼度が一定のため，フィラー信頼度のしきい値を大きくすると*FA*は減少するものの*FR*は増加する。また，許容値が小さくなるに従って正解と判定される区間が少なくなるため，*FA*と*FR*ともに増加する。

図6からはフィラー信頼度のしきい値及び許容値を変化させた場合の*FA*と*FR*の傾向は見られるものの，最適なフィラー信頼度のしきい値や許容値を求めるのは難しく，認識結果を確認して決める必要があると考えられる。今後，可読性と認識精度を考慮してそれらの最適値を決定する。

これらの結果から，今回開発した音声認識エンジンでフィラー検出ができることを確認した。フィラーの検出・除去例を，**図7**に示す。単語ごとにフィラーかどうかを判定しているため，可読性が求められるときには検出したフィラーを除去することができる。

4.3　話し言葉の音声認識精度

次に，フィラー検出機能を搭載した音声認識エンジンを用いて，音声認識精度を算出する。使用する音響モデルはフィラーラベルと音節ラベルを用いて学習を行った。また，音節をバランス良く学習するために，当社独自の学習セット構築法[6]を用いた。認識精度の尺度として，入力した文字列と認識された文字列のマッチングを取り，文字正解精度を算出した。ここでは，フィラー検出精度に関係なく，文字が正解と一致して

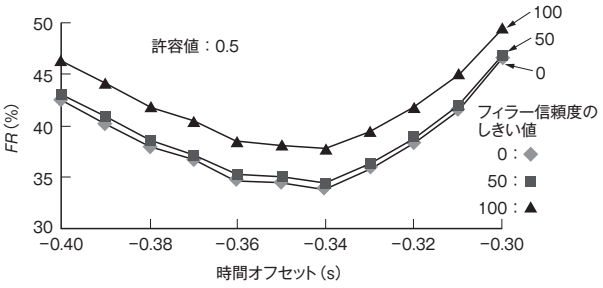


図4．時間オフセットとフィラー検出精度の関係 — どのフィラー信頼度のしきい値でも，時間オフセットが-0.34 sのときに*FR*がもっとも低くなっている。

Relationship between time offset and filler word detection performance

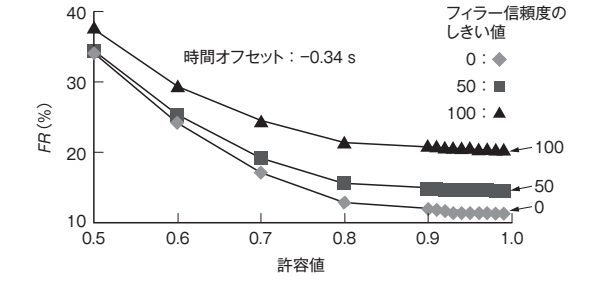


図5．許容値とフィラー検出精度の関係 — 許容値が大きくなると，*FR*は低下し0.8～0.9程度で収束している。

Relationship between tolerance and filler word detection performance

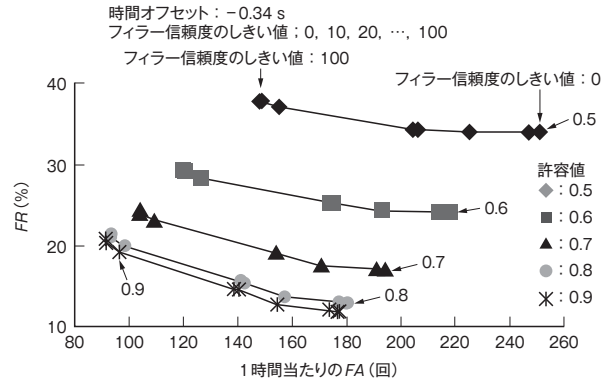


図6．許容値ごとのフィラー検出精度と1時間当たりの誤検出回数の関係 — 許容値が小さくなるに従って正解と判定される区間が少なくなるため，*FA*と*FR*ともに増加している。

Dependence of false acceptances (FA) per hour and false rejection (FR) rate on tolerance

例1：**と <F>** 札幌から，**えと <F>** 1時間ぐらい行くと
[フィラー除去] → 札幌から1時間ぐらい行くと

例2：**えーっと <F>** サークルは **えーと <F> まあ <F>** いろいろ入ってるんですけど
[フィラー除去] → サークルはいろいろ入ってるんですけど

例3：**えーと <F> その <F>** サークル活動の拠点が**越冬 <F>**
[フィラー除去] → サークル活動の拠点が

図7．フィラーの検出・除去例 — 今回開発した方式では，テキストからは通常の品詞と区別がつきにくい「その」や，別の漢字に変換された「越冬」なども，フィラーとして検出できている。

Examples of filler word detection and removal

表1．音声認識精度の評価結果
Results of evaluation of speech recognition performance

項　目	CSJテストセット3での文字正解精度（%）
フィラーラベル学習なし	89.56
フィラーラベル学習あり	90.21

いるかどうかで判定する。ここでの言語モデルはCSJに特化したものではなく，汎用のものを用いた。

評価結果を**表1**に示す。これより，フィラーを明示的に学習したほうが，高い認識精度を持っていることがわかる。この音響モデルと音声認識エンジンはフィラーを検出するだけでなく，明示的にフィラーを付与して学習することで，フィラーと通常の単語との境界をより高精度に学習できると考えられる。

5　あとがき

当社は，話し言葉音声認識の劣化要因となり，かつ音声認識結果から意味を理解することを妨げるフィラーを検出する，音声認識エンジンを開発した。開発した音声認識エンジンは，単語辞書に明示的にフィラーラベルを付与したフィラーを語彙として用意することなく，出力単語がフィラーか否かをデコーダー処理で判定する。

開発した方式により高精度にフィラーを検出でき，話し言葉音声認識において高精度な音声認識システムを実現した。その結果，会議やコールセンターなどの音声から可読性の高い音声認識結果を生成できるようになった。

今後，言いよどみに対してもデコーダー側で処理を行い，話し言葉の音声認識精度，及び音声認識結果の可読性を向上させ，話し言葉音声認識を用いた会議支援やコールセンター業務支援などに応用していく。

文　献

(1) 篠崎隆宏 他. 話し言葉音声の認識を目指して. 信学技報. **100**, 523, 2000, p.7－12.
(2) 那須　悠 他. LSTM-CTCを用いた音響イベント検出・除去音声認識システムの検討. 信学技報. **116**, 209, 2016, p.121－126.
(3) Graves, A. et al. "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks". ICML 2006 Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, PA, USA, 2006-06, ICML 2006, p.369－375.
(4) 永尾　学. “ノードのみで構成されるラティスを生成するWFSTに基づく音声認識デコーダ”. 日本音響学会2016年春季研究発表会講演論文集. 横浜, 2016-03, 日本音響学会. 2016, p.79－80.
(5) 前川喜久雄 他. 日本語話し言葉コーパスの設計. 音声研究, **4**, 2, 2000, p.51－61.
(6) Shinohara, Y. "A submodular optimization approach to sentence set selection". 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy, 2014-05, IEEE. 2014, p.4112－4115.

藤村　浩司　FUJIMURA Hiroshi
技術統括部 研究開発センター 知識メディアラボラトリー研究主務。音声認識技術の研究・開発に従事。日本音響学会会員。
Knowledge Media Lab.

益子　貴史　MASUKO Takashi, D.Eng.
技術統括部 研究開発センター 知識メディアラボラトリー主任研究員，博士（工学）。音声情報処理に関する研究・開発に従事。電子情報通信学会，日本音響学会，ISCA，IEEE会員。
Knowledge Media Lab.

永尾　　学　NAGAO Manabu
技術統括部 研究開発センター 知識メディアラボラトリー研究主務。音声認識技術の研究・開発に従事。日本音響学会会員。
Knowledge Media Lab.