割込み発話を認識できる音声認識システム

Barge-in Tolerant Speech Recognition System

鈴木　薫　　山本　幸一
■SUZUKI Kaoru　■YAMAMOTO Koichi

音声認識システムが応答音声を出力中に利用者が発話する割り込み発話では，マイクが収音する応答音声のエコーにじゃまされて利用者音声の認識が著しく困難になっている。

東芝は，このような課題に対応するため，エコーキャンセラを組み込んだ音声認識エンジンを開発し，マイク入力から応答音声のエコーを除去して，利用者の音声を正しく認識できるようにした。更に残留エコーを誤って認識しないよう，応答音声の強い周波数帯域を無視するように音声区間検出器（VAD：Voice Activity Detector）を改良した。このシステムを自動車内の音声対話に適用した結果，通常発話を従来と同程度に認識できるだけでなく，従来難しかった割込み発話も高い単語正解率で認識できることを確認した。また，改良されたVADの誤検出率が５％以内になる音量を調べ，実際の使用に耐えられることを確認した。

In a barge-in speech recognition system, when the user's utterance and the system prompt overlap, recognition of the user's speech may be interfered with by echoes of the system prompt.

Toshiba has developed a speech recognition engine equipped with an echo canceller to remove such echoes from the microphone input. However, as the echo canceller is not perfect, residual echoes remain in the output signal, and the voice activity detector (VAD) to detect the section appropriate to the user's speech may misdetect a residual echo. Consequently, we have improved the VAD to ignore the same frequency components as those used by the system prompt when calculating voice activity. The results of experiments assuming vehicle application showed that the word accuracy of this system for non-barge-in speech is the same as that of the conventional system, while high word accuracy for barge-in speech is also obtained. In addition, it was verified that the tolerable volume range of the VAD is sufficiently wide compared with the sound volume of the prompt that is actually used.

1　まえがき

利用者の音声を認識し，音声で応答する音声対話では，スピーカから出力された応答音声が空間を伝わってマイクに収音されるエコーが発生する。もしシステムが，エコーの混入した音声をそのまま認識しようとすると，エコーを利用者の音声と区別できなかったり，エコーの重畳した利用者の音声をうまく認識できないという問題が発生する。

そこで，従来の多くのシステムでは，応答音声の出力が終わってから音声の入力を始めることで，エコーを収音しない排他制御を行っている。しかし，排他制御は利用者が応答音声の終了を待つか，明示的な操作によりこれを中断させてからでないとシステムに向かって話すことができず使い勝手が悪い。

更に実際の場面では，このような中断操作の仕組みを備えていても，これを使うことなく利用者が話し始めてしまう割込み発話が発生する。例えば，利用者がシステムの質問に対してその音声が終わる前に思わず答えてしまう場合や，応答音声の終了を待ったつもりで最後の部分で声が重なってしまう場合などである。システムはこのような割込み発話を正しく認識できず，利用者は同じ内容を再度言い直す必要がある。

割込み発話ができて使い勝手の良い音声対話を実現するために，システムは応答音声出力とともに音声入力を開始し，この入力音声に混入するエコーの影響を取り除いて利用者の音声を認識できなければならない。しかし，音声入力の開始時期を早めれば，利用者音声以外の雑音の混入機会が増えることになり，システムには相応の頑健性が必要になる。

東芝は，この課題に対応するため，エコーキャンセラを組み込んだ音声認識エンジンを開発し，マイク入力から応答音声のエコーを除去して認識できるようにした。

ここでは，この音声認識システムの概要と特長について述べる。

2　割込み発話を許容する音声認識システム

2.1　音声認識システムのタイムシーケンスとブロック構成

今回開発した割込み発話を許容する音声認識システム（以下，提案システムと略記）のタイムシーケンスを**図1**に示す。

利用者がトークスイッチを押すと，システムは応答音声，つまり所定の入力を促すガイダンス音声とそれに続くビープ音をスピーカから出力する。応答音声はエコーになり，やや遅れて

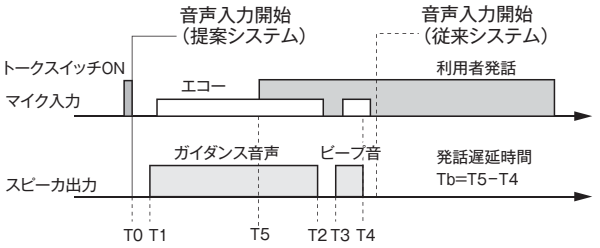

図1. 提案システムのタイムシーケンス ― 従来システムでは，排他制御のために応答音声出力完了後に音声入力を開始していたが，割込み発話を許容する提案システムでは応答音声の出力に先駆けて音声入力が開始される。

Time sequence of newly developed system

マイクに到達する。このとき，応答音声終了時刻T4から利用者発話開始時刻T5までの時間Tbを発話遅延時間と呼ぶことにする。利用者の発話はTb ≧ 0のとき割込みのない通常発話，Tb ＜ 0のとき割込み発話になる。

排他制御を行う従来システムでは，音声入力開始が応答音声終了時刻以降になるため，応答音声のエコーを取り込むことはないが，利用者発話の先頭部分も取り込まれない。一方提案システムでは，音声入力開始がトークスイッチの押された直後であるため，利用者発話の全域を取り込むことができるが，入力音声には応答音声のエコーが混入してしまう。

そこで提案システムでは，エコーキャンセラを内蔵した音声認識エンジンを新たに開発し導入することで，マイク入力からエコーを取り除いて利用者音声を認識できるようにする（**図2**）。更に，この音声認識エンジンでは，マイク後段と適応フィルタ前段に同じ特性を持つハイパスフィルタ（HPF）を入れることで，信号に混入する直流成分や背景雑音の低域成分を除去できるようにする。このHPFは，例えば車載応用では低い周波数にパワーの集中している走行雑音を弱める効果がある。

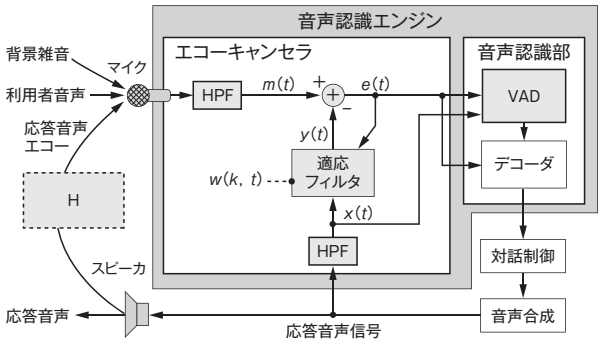

図2. 提案システムのブロック構成 ― 割込み発話を許容する提案システムは，音声認識部の前段にエコーキャンセラを備えた音声認識エンジンを持つ。

Block diagram of newly developed system

2.2 エコーキャンセラ

音声認識エンジンに内蔵されているエコーキャンセラは，スピーカからマイクまでの音の伝わりかた（伝達特性H）を適応フィルタによって模擬する。このフィルタによって模擬されたHを使ってスピーカへの信号（参照信号：図2の$x(t)$）を加工すれば，エコーとそっくりな信号（エコーレプリカ信号：図2の$y(t)$）を作ることができる。そして，これを入力信号（マイク入力信号：図2の$m(t)$）から引けばエコーの含まれない音声信号（出力信号：図2の$e(t)$）が得られる。

提案システムはこの$e(t)$を認識することで，エコーにじゃまされることなく利用者音声を認識でき，利用者はシステムが応答音声を出力中であっても中断操作なしで自由に発話できるようになる。

適応フィルタでは，N個のタップ係数wを使い，スピーカから出た音がどれくらいの遅れと強さでマイクに届くかを表現する。ここで，Hは時間とともに変化するので，wも遅れkと時刻tの関数として$w(k, t)$のように記述される。また，kの最大値を与えるNが大きいほど，フィルタは大きな遅れのエコーまで再現できる。この$w(k, t)$と$x(t)$を，式(1)のようにkが同じものどうしを掛けて足し合わせることで，$y(t)$を計算する。

$$y(t) = \sum_{k=0}^{N-1} \{ w(k,t) \times x(t-k) \} \qquad (1)$$

このとき，Hを模擬するタップ係数値を求めるために，勾配（こうばい）法を用いて$w(k, t)$を正しい値に漸近させる。今回は計算コスト，エコー消去能力，及び調整の容易さを考え合わせ，勾配法としてNLMS（Normalized Least Mean Squares）法を採用した[1), 2)]。NLMS法は$x(t)$と相関のある信号成分，すなわちエコー成分が$e(t)$の中で最小になるwを求めるように動作する。

2.3 参照信号を利用したVAD制御

2.3.1 利用者音声の音量に寛容なVAD特徴量

図2のVADは入力信号の各部分が音声か非音声かを判別する。これに続くデコーダはVADで音声と判別された区間の入力信号を認識する。認識結果は対話制御部で解釈されて対話が進行し，必要に応じて音声合成部からシステム応答音声が出力される。以上のようにして，提案システムは利用者と対話を行う。

VADは音声か非音声の判別に際して，正規化スペクトルエントロピー，平均SNR（Signal Noise Ratio），及びスペクトル間余弦値という三つの特徴量を使って入力信号の音声らしさを評価する[3)]。これらは既に報告されている特徴量[4), 5)]を雑音信号で正規化したものである。雑音信号は利用者発話と応答音声のない期間，例えば音声入力開始から応答音声出力までの間（図1のT0からT1の間）のマイク入力信号から推定される。

VADが使用するこれらの特徴量は，入力信号の絶対的な大きさなどを評価しないように設計されている。これは，VAD

が音量の大小にかかわらず様々な人の声を検出できるようにするためである。しかし，この特長が次に述べる残留エコーの問題を発生させる。

2.3.2　VADにおける残留エコーの問題と回避法

エコーキャンセラで使われる適応フィルタが正しいタップ係数値に近づくまでには時間が掛かる。そのため，フィルタの適応初期やエコーパスの変動直後には大きなエコーの消え残り（残留エコー）が発生してしまう。特に割込み発話では，応答音声出力開始から利用者の発話開始までにフィルタを適応させられる時間が十分長くとれないこともエコーが消え残る一因になっている。

一方前述したように，大きな声も小さな声も検出できるように設計されたVADは，システム応答音声が人の声の特徴を備えているため，わずかな残留エコーを音声と誤って検出してしまう可能性がある。そこで，$x(t)$ が使っている周波数帯域（主要周波数帯域）を前述の特徴量の計算から除外することで，応答音声の残留エコーに対して頑健で，かつ利用者音声の音量の違いに寛容なVADを実現する。

2.3.3　主要周波数帯域の選択と除外　$x(t)$ の細かい時間変動の影響を排除するために $x(t)$ のパワースペクトルを時間方向に平滑化し，以下の二つの基準をともに満たす周波数帯域を主要周波数帯域として選択する。

基準1　パワーの大きな周波数帯域を選択するために，平滑化後のパワースペクトルについて，そのときのスペクトル全体の平均パワーを所定係数倍した値をしきい値とし，この値よりもパワーの大きい周波数を選択する。

基準2　パワーの小さな周波数帯域，つまり残留エコーを生じる可能性の低い周波数帯域を除外するため，平滑化後のパワースペクトルについて，あらかじめ設定されたしきい値よりもパワーの小さい周波数を除外する。

これを模式的に表したものが**図3**である。基準1と基準2をともに満たす周波数成分を主要周波数帯域として選び出す。

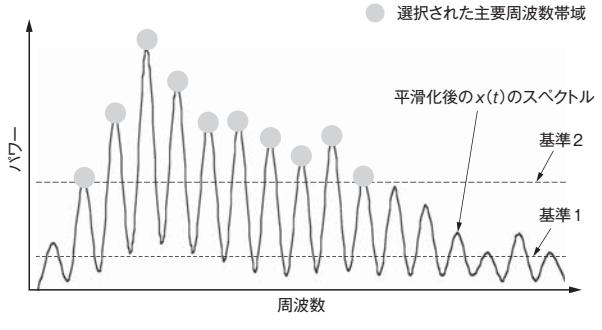

図3．主要周波数帯域の選択 ― $x(t)$ の平滑化後のパワースペクトルについて，基準1と基準2をともに満たす周波数成分を選び出し，その周波数を主要周波数帯域とする。

Selection of major frequency bands

VADでは，以上のようにして選択された主要周波数帯域を除いて前述の特徴量の値を計算し，残留エコーに影響されないようにして音声区間を検出する。

3　実車音声による評価実験

3.1　実験条件

前述の提案システムを自動車内の音声対話に適用することを想定し，**表1**に示す実験条件で，実際の車内で7名の話者に通常発話と割込み発話をそれぞれ4,200ずつ発話させた。発話間で性能の独立性を保つために，各発話の処理開始時にタップ係数の初期値をすべてゼロとした。

3.2　タップ数の探索

自動車内の残響時間は50 msと言われており，この時間をカバーする N は16 kHzサンプリングで800に相当する。そこで，N を800から100まで減らしたときの提案システムの単語正解率を比較した。その結果を**表2**に示す。

この実験によって，N を800付近にするよりも，300 ～ 400付近にするほうが認識性能は良いことが確認された。

3.3　従来システムと比較した単語認識性能

応答音声終了後に音声の入力を開始する排他制御を行う従来システムA，排他制御を行わなかった場合の従来システムB，

表1．自動車内の実験条件

Experimental conditions in vehicle

項　目	内　容
走行条件	アイドリング，市街地，高速道路
車種	スタイルの異なる2車種（A，B）
マイク	マップランプ位置に2種類
応答音声	長短2種類 話者の聞きやすい音量に調節
話者	男女7名 男性2名（車両A） 男性3名，女性2名（車両B）
発話内容	50都市名（辞書は100都市名）
発話数	通常発話　：全4,200発話 割込み発話：全4,200発話 細かいタイミングは話者に一任

表2．タップ数別の単語正解率

Results of word accuracy according to tap count (*N*)

N	単語正解率（%）	
	通常発話	割込み発話
800	95.74	92.48
500	96.00	93.02
400	96.05	93.07
300	95.98	93.33
200	95.60	93.00
150	95.29	92.29
100	95.10	91.24

表3. システム別の単語正解率
Word accuracy of each system

		従来システムA	従来システムB	中間システム	提案システム
音声入力		排他制御	排他制御せず		
単語正解率 (%)	通常発話	95.31	14.43	63.50	96.05
	割込み発話	対象外	52.26	73.52	93.07

排他制御を行わずエコーキャンセラ (N=400) だけを導入した中間システム，及びVAD制御まで導入した提案システムの単語正解率を比較した。その結果を**表3**に示す。

従来システムAと提案システムの比較から，通常発話に対する単語正解率は，排他制御を行う従来システムAの95.31 %に対し，排他制御を行わない提案システムでも96.05 %と同等の性能であることが示された。提案システムでは通常発話に対する不利が予想されたが，従来システムとそん色のない性能を示している。更に，排他制御時には認識対象外だった割込み発話に対しても，提案システムは93.07 %の高い性能を示すことが確認された。ここで，提案システムでの通常発話と割込み発話の単語正解率に約3ポイントの差が生じているのは，残留エコーによる利用者音声のひずみが原因と考えられる。

また，従来システムB，中間システム，提案システムの比較から，エコーキャンセラの導入によって単語正解率が改善するものの十分ではなく，高い認識性能を達成するためにVAD制御の導入が有効であることも確認できた。

3.4 VADの応答音声音量への耐性評価

また，この実験から，7人の話者全員が応答音声をストレスなく聴ける音量はおおむね88 dB以下であると判明した。そこで，この音量を常用範囲と呼ぶことにする。

これに対して，提案システムのVADが応答音声を誤って検出する率（VAD誤検出率）が5 %以内に収まる音量を耐用範囲と呼ぶことにする。Nを400とし，**表4**に示す実験条件で，市街地と高速道路を走行中の車内で音量を変えながら応答音声を出力し，VADがこれを誤検出する率を計測して耐用範囲を求めた。実験の結果，音量耐用範囲は，走行条件が市街地の場合89.4 dB以下，高速道路の場合93.7 dB以下になった。

表4. VAD誤検出率を計測する実験条件
Experimental conditions for measuring VAD detection error rate

項　目	内　容
走行条件	市街地，高速道路（80 km/h以上で巡航）
車種	1車種
マイク	マップランプ位置に1種類
応答音声	声質の異なる2種類
システム発話数	全433発話（音量ごとに20～50発話） 出力音量を数段階に変えて調査

いずれの走行条件でも耐用範囲の上限が88 dBを超えている。すなわち，通常使用される程度の音量であればVADは耐用範囲内で動作し，その誤検出率は5 %までである。

なお，市街地よりも高速道路での耐用範囲が広いことも判明した。これは応答音声の音量を上げざるをえない走行雑音の増大が，逆に残留エコーを目だたなくさせる結果になり，誤検出の防止に有利に作用していると考えられる。

4 あとがき

ここでは，割込み発話を許容する音声認識システムと，このシステムのために開発した音声認識エンジンについて述べた。この音声認識エンジンはシステム応答音声のエコーを除去するエコーキャンセラと，そこで消え残ったエコーに対して頑健なVADを備えている。車載応用を想定した実車収録音声に対する性能評価によって，開発した音声認識エンジンが通常発話を従来と同程度に認識できるだけでなく，従来難しかった割込み発話も高い正解率で認識できることを確認した。

また，VADの耐えられる応答音声の音量を調査し，通常使われる音量の範囲内であれば，応答音声を誤って検出する率は5 %以内であることを明らかにした。今後はいっそうの性能向上を図ることで，高い実用性を確保する。

文　献

(1) 大賀寿郎，ほか. 音響システムとディジタル処理. 東京，電子情報通信学会，1995，261p.
(2) Benesty, J., et al. Advances in Network and Acoustic Echo Cancellation. Germany, Springer, 2001, 222p.
(3) 鈴木　薫，ほか. “割り込み発話に頑健な音声認識エンジンの開発”. 日本音響学会 2010年秋季研究報告会. 大阪，2010-09，日本音響学会. 講演番号2-9-2.
(4) Ding, H., et al. "Comparative Evaluation of Different Methods for Voice Activity Detection". Proc. of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008). Brisbane, Australia, 2008-09, International Speech Communication Association. p.22－26. (CD-ROM).
(5) 山本幸一，ほか. 雑音にロバストな音声と非音声の判別技術. 東芝レビュー. **64**，12，2009，p.41－44.

鈴木　　薫　SUZUKI Kaoru
研究開発センター 知識メディアラボラトリー研究主務。
文字・図面・画像認識，及び音響信号処理技術の研究・開発に従事。情報処理学会，システム制御情報学会会員。
Knowledge Media Lab.

山本　幸一　YAMAMOTO Koichi
研究開発センター 知的財産担当主務。
音声認識技術の研究・開発を経て，現在，知的財産担当業務に従事。
Knowledge Media Lab.