背景音声雑音に頑健な特定話者発話区間検出の検討*

☆伊東直哉 (東大・情報理工), 松田繁樹, 柏岡秀紀 (NICT),
辻野孝輔 (NTT ドコモ), 嵯峨山茂樹 (東大・情報理工)

1 はじめに

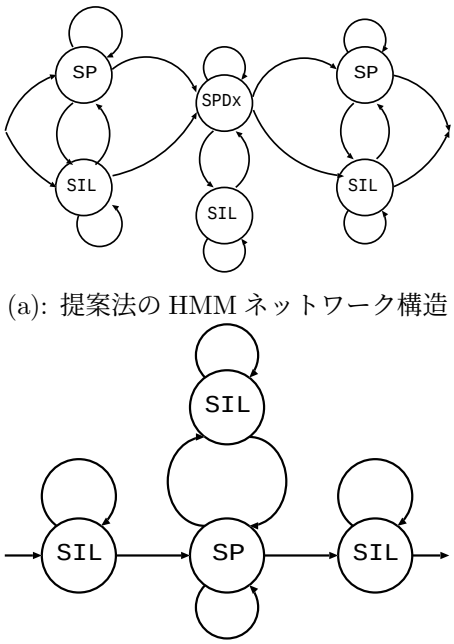
実環境下での利用を想定した音声認識システムは, 背景雑音に対して高い頑健性が求められる. 雑音に対する頑健性の改善のため, 様々な雑音抑圧手法 [1, 2] や, 発話区間検出手法 [3], 複数の音響モデルを用いて認識する手法 [4] など数多くの手法が提案されてきた.

独立行政法人情報通信研究機構 (NICT) では, ネットワーク型多言語音声翻訳アプリケーション「VoiceTra」のサービスを通して実利用音声の収集及び分析を進めている. 実利用音声を用いた評価実験等から得られた知見として, ユーザ以外の人, 例えば飲食店で隣に座った別のお客の声や, 友人の声, 子供の「わたしにもつかわせてー」など人の声によって, 単語挿入誤りが顕著に増加することを確認している.

音声強調などの雑音抑圧手法を用いて前述のような人の声を含む背景雑音 (背景音声雑音; Background Speech Noise： BSN) を抑圧することは, 雑音に含まれる音声が強調されてしまうため, 一般に困難である. そこで本稿では, ユーザが発話した音声区間のみを切り出す特定話者発話区間検出法について検討を行った. 従来の混合ガウス分布モデル (Gaussian Mixture Model： GMM) を基礎とした発話区間検出手法 (Voice Activity Ditection: VAD) では, 一般に, 音声 GMM と雑音 GMM の 2 つの GMM の間の尤度比が用いられる. 提案法では, それに加えユーザの音声に適応された特定話者音声 GMM を併用することにより, ユーザの音声のみを切り出す特定話者発話区間検出が実現できると考えられる. 本稿で提案する VAD を用いることにより, ユーザ以外の音声区間 (前述の隣の客や友人など) を対象から外すことが可能となり, 誤った単語の挿入による性能劣化を抑えることができると考えられる.

2 特定話者発話区間検出法

Fig. 1(a) に提案する VAD のための隠れマルコフモデル (Hidden Markov Model: HMM) の構造を示す. 図中の SIL は雑音 GMM, SP は不特定話者音声 GMM, SPDx は特定話者音声 GMM を表す. 図に示すように, 提案する特定話者 VAD は SPDx が SP に挟まれた構造を持ち, 入力音声を Viterbi アラインメントすることにより,SPDx にアラインメントされた音声区間がユーザの発話した音声区間として判定される. 一方、BSN に含まれるユーザ以外の発話は, SP の状態にアラインメントされると考えられる. 提案法で用いる特定話者音声 GMM (SPDx) は, 不特定話者音声 GMM (SP) に対してマルチクラス最尤線形回帰 (Maximum Likelihood Linear Regression: MLLR) を用いた話者適応により推定される.



(a): 提案法の HMM ネットワーク構造

(b): 従来法の HMM ネットワーク構造

Fig. 1 VAD のための HMM ネットワーク構造: (a): 提案法, (b): 従来法

Fig. 1(b) に, 音声 GMM (SP) と雑音 GMM (SIL) のみから構成された従来型 VAD の構造を示す. 入力音声に BSN が含まれていた場合, 切り出し対象となる音声区間だけでなく BSN のいずれも SP にアラインメントされるため, ユーザ以外が発声した音声区間も発話区間として検出されると考えられる.

3 特定話者発話区間検出法の実験

3.1 実験条件

提案法による特定話者に対する VAD の有効性を示すため, 発話区間検出実験を行った. 音声データベー

*Study of speaker-dependent voice activity ditection robust to background speech noise by ITÔ Naoya(The University of Tokyo), MATSUDA Shigeki, KASHIOKA Hideki(NICT), TSUJINO Kosuke(NTT docomo), SAGAYAMA Shigeki(The University of Tokyo)

Table 1　用意するデータセット

データセット	学習用	適応用	評価用	
			w/o BSN	w/ BSN
人数	400	40(特定話者)		
発話数	約 70	200	100	
重畳雑音	15 種類		5 種類	

スは，学習用に音素バランス文データベース (TRA-BLA) と旅行会話文データベース (TRA) を，評価用に旅行会話基本表現集 (BTEC) を用いた．評価用データセットは，BSN 無しのもの (w/o BSN, ただし音声雑音でない雑音を含む) と有りのもの (w/ BSN) を用意した．

雑音は，車や電車など 20 種類の環境雑音の中から，15 種類を学習用及び適応用データセットに重畳し，残りの 5 種類を評価用データセット (w/o BSN,w/ BSN) に重畳した．SNR は 15 dB, 20 dB, 25 dB, 30 dB の 4 種類とした．BSN は，切り出し対象発話との SNR が 12 dB になるように重畳した．

音響特徴量は，12 次元の MFCC (Mel Frequency Cepstrum Coefficient) 及び，時間微分特徴量である 12 次元 ΔMFCC，Δ パワーの計 25 次元である．サンプリング周波数 16 kHz, 分析窓長 20 ms, 分析周期 10 ms で分析を行った．

VAD の評価には，False Rejection Rate (FRR), False Acceptance Rate (FAR) を用いた．

$$FRR = \frac{N_{FR}}{N_s} \times 100[\%] \quad (1)$$

$$FAR = \frac{N_{FA}}{N_{ns}} \times 100[\%] \quad (2)$$

N_s は音声フレーム数,N_{FR} は音声を非音声として検出したフレーム数，N_{ns} は非音声フレーム数，N_{FA} は非音声を音声として検出したフレーム数である．

マルチクラス MLLR を用いた話者適応では，その行列変換のクラス数 C が $1, 8, 32$ の場合について実験を行った．なお，比較のために,Fig. 1(b) で示した従来法の実験も行った．

3.2　実験結果

特定話者発話区間検出法の実験結果を Fig. 2 に示す．BSN を含むテストセットに対する評価において，提案法の FAR 及び FRR の値は，従来法より低下した．これは，従来法では BSN の区間が音声区間として判定されていたのに対し，提案法では，不特定話者 GMM (SP) にアラインメントされることで非音声区間として判定されたためである．

また，C の値を大きくするほど値が下がる傾向が見られ，適応のクラス数はある程度大きいものが必要であることを示している．

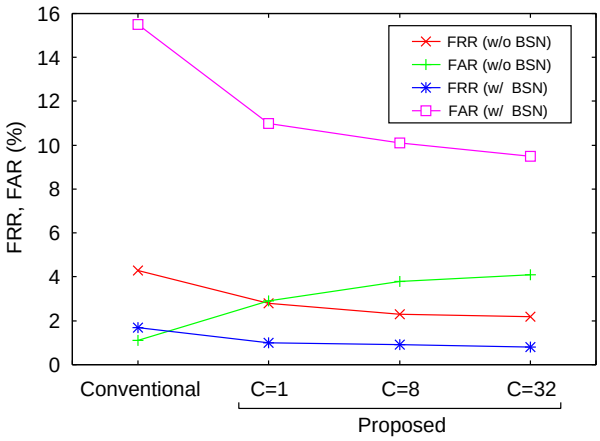


Fig. 2　特定話者発話区間検出の実験結果

BSN なしのデータのセットに対する評価においては，提案法の FRR の値は従来法より低下するが FAR の値は従来法より若干上昇している．

4　おわりに

本稿では，背景音声雑音に頑健な特定話者発話区間検出法を提案し，実験によって FAR,FRR の値が下がることを確認した．

今後は，特定話者発話区間検出法の音声認識における有効性を示すために，音声認識性能での比較実験を行う予定である．

謝辞　本稿は，NICT における夏期インターンシップによる成果である．また，有益なご助言をいただいた東京大学齋藤大輔助教，亀岡弘和客員准教授，国立情報学研究所小野順貴准教授に謝意を表する．

参考文献

[1] Segura, J. C., et al., “Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA 2 database and tasks”, Proc. EuroSpeech ’01, Vol.1, 221-224, 2001.

[2] M. Fujimoto, et al., “A Non-stationary Noise Suppression Method Based on Particle Filtering and Polyak Averaging”, IEICE Transactions on ED, Special Section of Statistical Modeling for Speech Processing, vol. E89-D, no.3, pp.922-930, 2006.

[3] 藤本 et al., “確率モデルに基づく音声区間検出と雑音抑圧の統合の検討”, 音講論 (春), pp.27-30, 2008.

[4] S. Matsuda, et al., “ATR Parallel Decoding Based Speech Recognition System Robust to Noise and Speaking Styles”, IEICE Transactions on ED, vol. E89-D, no.3, pp.989-997, 2006.