背景音声雑音に頑健な主発話区間検出に関する検討*

☆尾崎洋輔, 柏木陽佑, 齋藤大輔, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

音声認識をアプリケーションとして利用する場合に考慮すべき点として耐雑音性が挙げられる．雑音下では認識精度が大幅に劣化してしまうので，頑健な音声認識を実現するためには，(1) 認識したい発話が存在する区間を推定する発話区間検出 (VAD; Voice Activity Detection)，(2) 特徴量の段階でノイズを抑制する，あるいは音声特徴を強調する，といったアプローチが有効である事が知られている．しかし一般的な VAD や雑音抑制などでは，背景雑音が非音声である事を前提として音声と非音声の音響的差異に基づいて検出を行っていたり，雑音に定常性を仮定するなど制約がある場合が多い．そのため，雑音自体が “音声らしい特徴” を持ち合わせている場合には（明瞭な人の声による雑音を本稿では音声雑音と呼ぶ），これらの手法が上手く働かない事が考えられる．特に VAD を行う際に音声雑音が混入すると，認識対象の話者の音声（以下，主発話区間と呼ぶ）に加えて音声雑音区間も含まれた音声セグメントが得られる事になる．その中でも主発話区間外に挿入される音声雑音は，その区間の主要な成分としてその後の処理に大きな影響を与えてしまう．そこで，本研究ではこの音声セグメントに対して，さらなる前処理として主発話区間検出を行い，音声雑音に頑健なシステムの構築を目指す．

主発話区間検出に有用な特徴としては (a) 話者性の違い，(b) 空間伝達特性の差異，(c) 音声のエネルギー，(d) タスクから想定される発話位置のようなメタ情報，などが挙げられる．この中でどのような特徴を用いるのかは想定されるタスクに依存する．そこで本研究ではマイクに向かって音声入力を行うというタスク想定を行い，以下のような仮定を置く．

1. 入力信号は予め発話単位 (主発話+雑音) にセグメントされており，必ず主話者が 1 人だけ存在する
2. 主発話区間は音声セグメントの大部分を占める
3. 主発話区間のパワーは雑音区間のパワーより十分に大きい
4. 主話者に関する情報は未知である

1，2 はポーズやシステムとの対話情報を用いて入力信号からおおまかに発話区間を切り出した結果であり，3 は実環境におけるマイクと話者の距離を考慮している．本研究では，この中でも特に (a) 話者性の違いに着目して，背景音声雑音に頑健な主発話区間検出を試みる．

2 関連研究

2.1 VAD

VAD に一般的に用いられる特徴量としては入力信号のパワー，スペクトル，零交差数，音声モデルと雑音モデルの尤度比，帯域間のスペクトル密度比，およびそれらの動的特徴や長時間特徴などがある [1]．ただし，これらの多くは人の声とそれ以外の雑音の特徴の差異に着眼しており，主発話と音声雑音の分離に関しては効果が薄いと考えられる．識別手法としては，上述した特徴量に対して閾値を設けるというのが最も基本的な手法となる．また，HMM (Hidden Markov Model) のような生成モデルを用いて尤度最大基準で区間検出するといった手法も存在する．近年では SVM (Support Vector Machine) や Neural Network のような識別モデルを用いて音声/非音声を識別する研究もされており [2, 3]，音声認識などで一般的に使用される MFCC (Mel-Fequency Cepstrum Coefficients) とその派生特徴のみを SVM の入力に用いて高精度の VAD を実現している研究もある [4]．識別モデルの場合，様々な素性を統合的に扱う事が出来るという利点がある．

2.2 音声雑音に関する検討

音声雑音を考慮した VAD を行った研究としては，[5] や [6] が挙げられる．[5] は，予め学習された主話者の GMM (Gaussian Mixture Model) と不特定話者による UBM (Universal Background Model) を用いて主発話区間検出を実現しており，これは音声信号の話者性に着眼した手法と言える．また，ここでは発話状況がある程度絞られた状況である事を想定し，HMM を用いて入力信号の時間構造も含めてモデル化を行っている．ただし，この手法では予め主話者の音声を学習する必要があるという実用上の制約が存在する．他にも [6] では，音声の空間伝達特徴を手掛かりとした処理を行っている．これは主話者の声とその他の話者による声では直接音と反射音の伝達特性に明確な差があるという考えに基づくものであるが，現時点では効果のある場面が限られており課題も多い．

3 提案手法

3.1 提案手法の概要

今回の研究は話者非依存タスクとなので，“入力話者が誰” で “他話者とどのような違いが声質に現れるのか” という点に関しては，予め学習しておく事が出来ない．そこで話者性という話者依存の特徴を，如何にして話者非依存の “主話者らしさ” という特徴に落とし込むかが重要となる．本研究では処理を 2 段に分け，1 段目でおおまかな主話者区間検出を行い，その結果を用いて 2 段目で “主話者らしさ” のモデル化・スコア算出を行う．提案手法の概要を Fig. 1 に示す．初段での処理手順は以下のようになる．

---

*Voice Activity Detection for the Main Speaker under Background Speech Noise by Y. Ozaki, Y.Kashiwagi, D. Saito, N. Minematsu and K. Hirose (The University of Tokyo)

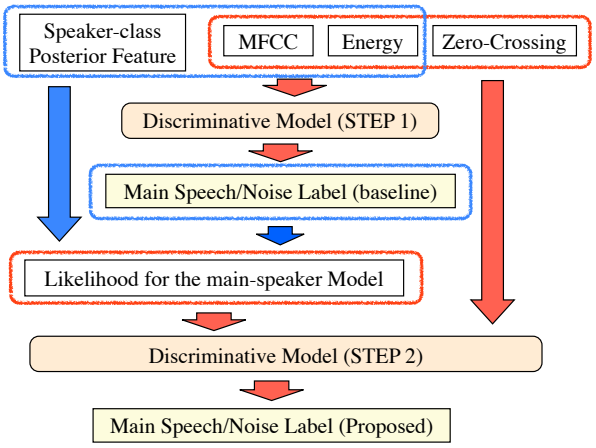

Fig. 1　提案手法の概要

1. MFCC，エネルギー，零交差数などの特徴を抽出
2. 1 の特徴を用いて主発話区間を検出

初段の区間検出には一般的に用いられる特徴を用いてるので，ここまでの処理がベースラインとなる．後段では，ここで得られた識別ラベルとその信頼度[1]を用いて，音声セグメント毎に異なる主話者の特徴をモデル化する．

1. 信頼度を利用して “主話者らしさ” をガウス分布でモデル化
2. 1 で得られたモデルに対するフレーム毎の尤度スコアを計算
3. 2 で得られたフレーム特徴を 1 の特徴に連結する
4. 3 の特徴量を用いて再度 VAD ラベルをフレーム毎に推定（提案手法）

2 段目の 1–2 の操作は話者依存特徴を取り扱う為に発話セグメント内で完結しており，逆に話者非依存特徴を用いる 1 段目と 2 段目の主発話区間検出では，学習セット全体を用いて識別器を学習している．

3.2　“主話者らしさ” を表すフレーム特徴

前述したように後段の処理では，フレーム毎の信頼度スコアを入力として利用する．この信頼度スコアを用いて以下のように重み付き平均・分散を算出する事で，発話毎の主話者モデルを単一のガウス分布でモデル化する．

$$w'_t = \begin{cases} w_t & (w_t > 0) \\ 0 & (w_t \leq 0) \end{cases} \quad (1)$$

$$W = \sum_{t=1}^{T} w'_t \quad (2)$$

$$\mu = \frac{1}{W}\sum_{t=1}^{T} w'_t X_t \quad (3)$$

$$\Sigma = \frac{1}{W}\sum_{t=1}^{T} w'_t (X_t-\mu)(X_t-\mu)^{\mathrm{T}} \quad (4)$$

[1]識別器から得られる閾値処理をする前のスコアや事後確率．実装によって異なる．

ここで，w_t は t 番目のフレームの信頼度スコアの初期値であり，$X_t(t=1,2,...,T)$ は音声セグメントのフレーム特徴系列である．今回は実装において，スコアが正負の値を取る線形識別器を用いたので，式 (1) で信頼度スコアが負であるフレームをカットしている．また，正規化を行うためにフレームの重みを等価とした UBM を作成する．

$$\mu_{\mathrm{ubm}} = \frac{1}{T}\sum_{t=1}^{T} X_t \quad (5)$$

$$\Sigma_{\mathrm{ubm}} = \frac{1}{T}\sum_{t=1}^{T} (X_t-\mu_{\mathrm{ubm}})(X_t-\mu_{\mathrm{ubm}})^{\mathrm{T}} \quad (6)$$

これらの分布に対する尤度を基に，主発話クラスについての事後確率を 2 クラス間のロジスティック回帰で近似を行い，識別関数に相当する特徴を設計した．

$$A(X_t) = \frac{\mathcal{N}(X_t\,;\,\mu,\Sigma)}{\mathcal{N}(X_t\,;\,\mu,\Sigma)+\mathcal{N}(X_t\,;\,\mu_{\mathrm{ubm}},\Sigma_{\mathrm{ubm}})} \quad (7)$$

ただし，ここで主話者クラスと UBM クラスの事前情報は未知なので，それぞれのクラスに対する事前分布を等価であると仮定している．こうして得られた特徴 $A(X_t)$ を話者非依存の “主話者らしさ” を表す特徴として主発話区間検出に用いる．

3.3　話者クラス事後確率特徴

上記のプロセスでは話者性に着目して “主話者らしさ” を表すフレーム特徴を定義した．しかし，MFCC などの特徴では長期間で見ると話者の特性が表れるが，短期間で見ると発話内容の影響が大きく単一のガウス分布では表現し切れないといった問題も考えられる．そこで，発話内容に左右されない（されにくい）ような話者性を表す特徴を新たに定義し，前述の “主話者らしさ” モデル化する際に，音響特徴量に連結して使用する．近年の音声認識でその有用性が示されている TANDEM アプローチ[2][7] を，音素クラスではなく予めテンプレートとして用意された話者クラスに対して適用し，話者性に関するフレーム特徴として “主話者らしさ” をモデリングする際に用いる．具体的な手法としては以下のようになる．

1. 入力信号から一般的な音響特徴量（今回は MFCC）を抽出する
2. 話者クラスに対する事後確率ベクトルをフレーム毎に計算
3. 2 で求めた事後確率を対数に変換
4. 主成分分析を適用して次元圧縮

ある程度の数の話者クラスを用意すれば，入力話者（主話者，背景話者）に近い話者性を持ったクラスがいくつか存在し，それぞれの話者クラスに対する事後確率が入力話者の特徴となると考えられる．こうして得られる特徴を本稿では話者クラス事後確率特徴と呼ぶ．Fig. 2 に話者クラス事後確率特徴をカラープロットしたものを示す．このサンプルにおいては，

[2]`Neural Network` などを用いてフレーム毎に音素クラスに対する事後確率を求め，それの派生特徴を元の短時間特徴に連結する手法

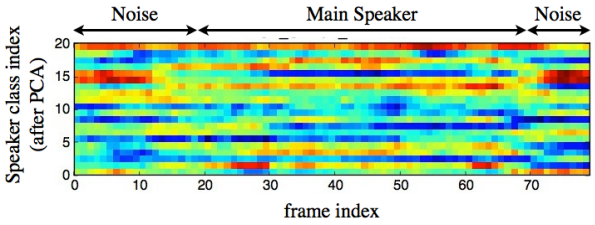

Fig. 2　話者クラス事後確率特徴 (SNR5)

縦軸は圧縮後の話者クラス事後確率ベクトルの次元である．ここに示したサンプルは SNR5 と音声雑音が大きいが，区間によって話者の特徴が表れている事が見て取れる．しかし話者クラス事後確率特徴が表す話者性は話者依存特徴なので，前項の “主話者らしさ” をモデル化する際に利用して話者非依存特徴に変換する．

4　実験

今回は 2 種類の実験を行った．

(I) “主話者らしさ” を求める際に使用する特徴量を数パターン変化させて実験

(II) ベースラインを信頼度スコアの初期値と見なし，主発話区間検出によって信頼度スコアを更新しながら提案手法における後段の処理を繰り返し適用

4.1　実験条件

(A) データベース

実験には東北大・松下単語音声データベース [8] を用いた．東北大・松下単語音声データベースは男女 30 人ずつ合計 60 人の話者による孤立単語 212 語彙の読み上げ音声であり，実験ではこの中から男女 15 人ずつの 30 人分の約 6300 発声分を学習用セットとして用いて，残りをテストセットとして用いる．また，本データベースはクリーン環境で収録されているため，音声雑音として JNAS (Japanese Newspaper Article Sentences) [9] の一部[3]を音声雑音として付与して雑音データベースとした．音声と雑音の比である SNR (Signal to Noise Ratio) はそれぞれ 5, 10, 15 として，これに元の音声を加えた 4 種類の環境下で実験を行う．

データベースの正解 VAD ラベルは既存の音素 HMM モデルを用いたモデルアライメントをクリーン音声に適用することで作成する．音響モデルには CSRC [10] の 16 混合の monophone モデルを使用し，/silE/，/silB/，/sp/，/q/の区間のラベルを雑音，それ以外の音素区間を主発話区間とした．

(B) ベースライン

本研究においては，提案手法における初段の識別結果が主発話区間検出のベースラインとなる．特徴量には Table 1 の条件で抽出された MFCC にフレーム毎の対数エネルギーと零交差数を連結して使用した．ベースライン，提案手法共に識別器の実装には機械学習ライブラリである oll (Online Learning Library) [11] の PA2 (Passive Agressive [12]) を用いた．

[3]東北大松下データベースの学習セットとテストセットのそれぞれに男女 5 名ずつの計 20 名分の新聞読み上げ音声を用いた．

Table 1　音響分析条件

サンプリング周波数	16 bit / 16 kHz
窓幅＆シフト長	25 ms length / 10 ms shift
特徴量	MFCC (12dim) + Δ + ΔΔ

Table 2　DNN 学習条件

特徴量	MFCC (12dim) + Δ + ΔΔ + E + ΔE + ΔΔE (39 dim) × 11 フレーム (前後 5 フレーム)
入力特徴次元	429
隠れ層	3 層
隠れ層のノード数	1024
識別クラス	60 名の話者（JNAS から）

(C) 話者クラス事後確率

話者クラスに対する事後確率の計算には DNN (Deep Neural Network) を用いており実装には Kaldi [13] を用いた．詳細な学習条件は Table 2 に示す．使用した話者クラスは 60 で，次元圧縮後の話者クラス事後確率の次元数を 20 した．

(D) 主話者らしさを表す特徴量

実験 (I) では “主話者らしさ” をモデル化する際に使用する特徴量を，MFCC : MFCC+エネルギー，POS : 話者クラス事後確率のみ，MFCC-POS : MFCC+エネルギー話者クラス事後確率，のように 3 種類の場合で実験を行った．また，ここで “主話者らしさ” は対角共分散ガウス分布でモデル化を行っている．こうして得られた特徴を，フレーム毎に前後 5 フレーム分をまとめてベースラインの特徴量セットに連結する事により，提案手法の特徴量セットとする．

(E) 評価方法

フレーム識別の評価は VAD の評価で一般的に用いられる FRR (False Rejection Rate) と FAR (False Acceptance Rate) で行う．今回の場合，FRR は主発話フレームを雑音区間と判別した割合であり，FAR は雑音フレームを主発話区間と判別した割合となる．これらはトレードオフの関係なので，2 次元上のグラフにプロットした ROC (Receiver Operating Characteristics) カーブを見る必要がある．これを定量的に判断するには，各軸と ROC カーブに囲まれた区間の面積である AUC (Area Under the Curve) を計算する事が有効である．本研究では主に AUC を基に VAD の精度を検証する．

4.2　実験結果

実験 (I)

提案手法のフレームワークと話者クラス事後確率の効果を確かめるために，MFCC，POS，MFCC-POS の 3 種類の実験を行って AUC を算出した．clean 学習した場合の結果を Fig. 3 に，multi 学習 (clean, SNR15, SNR10, SNR5) した場合の結果を Fig. 4 に示す．

雑音の無い clean 環境や雑音の少ない SNR15 の場合には大きな改善は得られなかったが，雑音の大きい SNR10，SNR5 の場合にはいずれの場合も提案手法が大きく上回る事となった．特に POS の改善度が最も高く，AUC の削減率で見ると平均で 33.7 %，また最も雑音の大きい SNR5 の場合には 44.9 %と大き

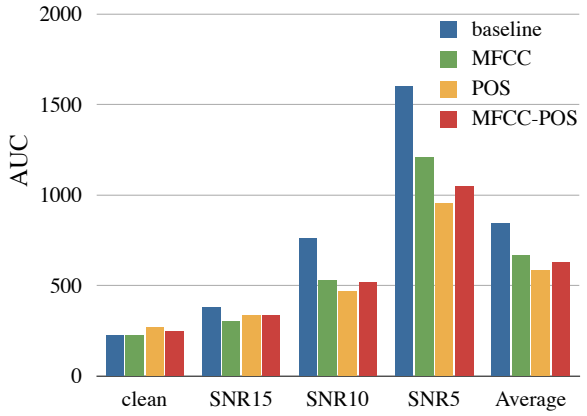

Fig. 3　実験 (I) clean 学習での AUC

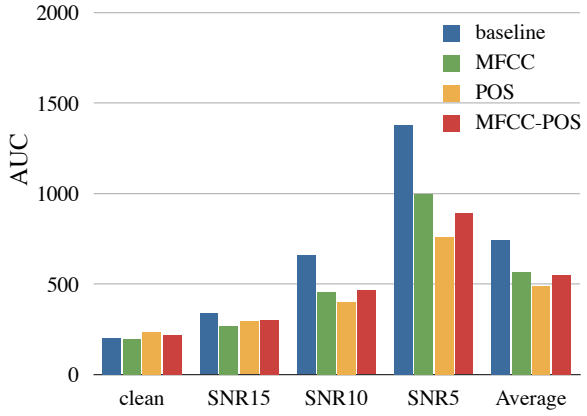

Fig. 4　実験 (I) multi 学習での AUC

な改善が見られた．また，識別器を clean 学習した場合でも，かなりの誤り改善効果が得られている事から，新たに追加した特徴が音声雑音に頑健であり，また雑音の大きさに左右されにくい事が分かる．

実験 (II)

実験 (I) では “主話者らしさ” の特徴量を求める際に，ベースラインの識別結果を信頼度スコアとして用いていた．しかし，ここで提案手法を用いて求めたより精度の高い主発話区間検出ラベルの信頼度を初期値と考え，(i)2 段目の識別器による主発話区間検出と (ii)“主話者らしさ” を表す特徴量の抽出，を繰り返し交互に行う事により主発話区間検出の精度向上が期待できる．そこで，実験 (I) で最も精度の高かった POS，multi 学習の条件で，1 回から 5 回まで繰り返し主発話区間検出実験を行った．結果を Fig. 5 に示す．oracle はベースラインにおいて clean 条件で算出された信頼度スコアを用いた場合の結果であり，信頼度スコアの質の向上による提案手法の精度の限界値と言える．繰り返しの回数が増えるに従って徐々に oracle に近付いていく事が予想されたが，実際には 1 度目で大きく改善した以降は，殆ど改善する事は無かった．これは oracle と比べてみると分かるが，1 度目の繰り返しで既に oracle に近い精度となっており，これ以降は大きな精度向上は期待できない．

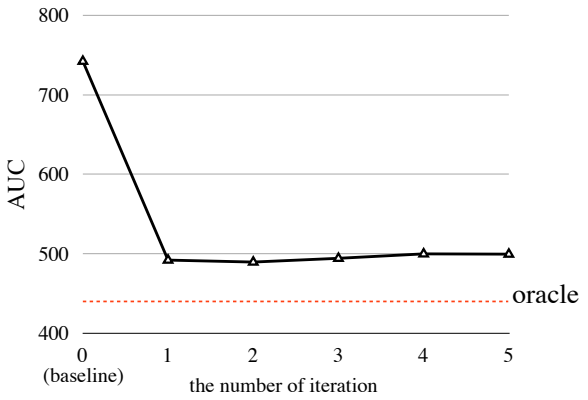

Fig. 5　実験 (II) 繰り返し学習を行った際の AUC (Average)

5　まとめと今後の展望

本研究では音声雑音下での主話者発話区間検出において，従来使用されていた音響特徴量に新たに話者性に関する処理を加える事によって，精度向上を目指した．話者性を表す特徴として話者クラス事後確率を新たに定義し，これを用いて音声セグメント毎の “主話者らしさ” をモデリングする事によって一定の成果を得る事が出来た．

今後の展望であるが，実験条件やそれぞれのステップにおける特徴量の取り扱いに関してはまだまだ改善すべき余地があるので，それらを厳密化する必要がある．話者クラス事後確率では話者数が増えて来ると計算量が増大する一方で，得られる特徴の各次元の持つ特徴が薄くなると考えられるので，話者クラスタリングを行うなどして適切に話者クラスを設定する必要がある．また始めに述べた通り，背景音声雑音に頑健な特徴は話者性の他にもいくつか存在するので，それらの特徴も併せて検討する必要がある．今回のタスクでは FAR と FRR を用いて提案手法の性能を検証したが，VAD は音声認識などにおける前処理の段階の話なので，その結果がアプリケーションなどにどのような影響を及ぼすかが重要になる．そのため VAD 単体の評価だけでなく，VAD の結果を音声認識などの周辺技術と組み合わせた場合についても検証する必要がある．また，フレーム単位での VAD においては，時間的な制約などを用いて識別結果を “ならす” 処理が必要となる．VAD の結果をアプリケーションに応用するにはこういった処理も必要となるので，こちらもいずれは検討したい．

参考文献

[1] 藤本, 電子情報通信学会誌, Vol. 95, No.8, 754-758, 2012.
[2] D. Enqing, et al. Signal Processing, vol.2, 1124-1127, 2002.
[3] X. Zhang, et al. IEEE TASL, vol. 21, No. 4, 697-710, 2013.
[4] T. Kinnunen, et al. SPECOM, 556-561 2007.
[5] S. Matsuda, et al. INTERSPEECH, 2626-2629, 2012.
[6] 鎌土 他, 音講論 (春), 145-146, 2010.
[7] H. Hermansky, et al. ICASSP, Vol. 3, 1635-1638, 2000.
[8] http://research.nii.ac.jp/src/TM212.html
[9] jnas, http://www.mibel.cs.tsukuba.ac.jp/_090624/jnas/
[10] CSRC, http://www.lang.astem.or.jp/CSRC/
[11] oll, http://code.google.com/p/oll/
[12] K. Crammer, et al. Journal of Machine Learning Research, 551-585, 2006.
[13] D. Povey, ASRU, 2011.