日本音響学会2011年秋季研究発表会講演論文集, pp.227-228 (2011)

サービスオペレーション推定における音声技術の導入*

☆竹原正矩，田村 哲嗣（岐阜大），天目隆平，蔵田武志（産総研），速水悟（岐阜大）

1　はじめに

本稿は，サービス産業（日本食レストラン，介護施設）におけるサービスオペレーション推定（Service-Operation Estimation; SOE）の概要と，音声区間検出やキーワードスポッティング等の音声技術の導入について紹介する．従来のサービス業務の分析・評価は従業員や顧客の主観に基づくものが多かった．筆者らは，サービスの提供者の行動を計測し，会計データ，業務データと併せて客観的な分析と評価を行う体系の構築を試みている [1]．本稿は，その体系の中で従業員の発話音声の音声区間検出（Voice Activity Detection; VAD）と，音声認識を用いて SOE に有用なキーワードの抽出に着目した．VAD から得られる発話量を SOE の特徴量の 1 つとして利用し，その有効性を検証した．

2　サービスオペレーション推定

2.1　サービスオペレーション推定の概要

SOE とは，サービス現場における従業員の作業内容（SO）を，行動データ（位置・方位・動作・音声），業務データ，会計データを用いて自動的に推定することである．SO とは，後述する日本食レストランの例では「挨拶・案内」「注文伺い」「配膳」等，複数の動作や発話，移動を組み合わせたものを指す．推定した SO は 2.3 節で紹介する可視化ツールで可視化し，業務中の SO や状況を再現する他，仕事量，負荷，作業時間など業務分析のための定量的なデータに変換する．

2.2　要素データの収集

計測するデータは，位置・方位・動作と発話音声である．また，同時に業務データ，会計データも収集する．これらのデータをまとめて要素データと呼ぶ．発話音声は，従業員に骨導音マイクを装着させて収録した音声である．骨導音マイクは気導音マイクと異なり，気導音による周囲の雑音の影響を受けにくい．これは，後述する音声処理の難易度を下げる他，収録時に顧客の発話内容に関するプライバシーに配慮する目的がある．

従業員の発話内容は，SO や行動に依存する場合が多く，音声認識によって推定したキーワードは SOE に有効であると考えられる．また，前処理である VAD によって求められた発話時刻や発話量といったデータは SOE の有効な特徴量に成り得る．

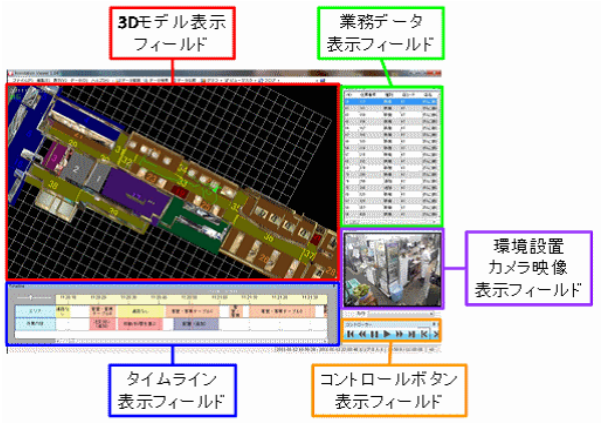

Fig. 1　A visualization example.

2.3　推定結果の可視化

要素データと SOE の結果は Fig. 1 に示す可視化ツールを用いて可視化を行う [1]．ユーザは要素データや従業員，時刻等を自由に選択して情報を表示させ，任意の時刻の店内の状況を知ることが可能である．

このツールは，経営者や従業員に対する QC（Quality Control）活動支援等に活用してもらうために，より良い分析・評価が可能な情報の提示と，インターフェースの改良を試みている．また，業務終了後すぐに分析・評価して欲しいという現場の要望もあるため，計測データの処理と SOE には速度やリアルタイム性が求められており，これらの検討も行っている．

3　骨導音データの音声処理

3.1　音声区間検出

骨導音データには最初に VAD を行い，マイク装着者の発話を抽出する．骨導音データは気導音データと比較して雑音や周囲の発話者のパワーは低減される．しかし，実際の収録では様々な作業においてマイク接触時の突発性ノイズが多く発生する．また，他者の発話もマイク装着者の身体を伝導し骨導音として重畳するため，音声認識の前処理として VAD が必要となる．本研究では，VAD によって発話抽出と同時に発話時刻，発話量を得て SOE や可視化の 1 データとして用いている．

骨導音マイクは気導音マイクに比べ，発話者の発話パワーが低周波成分に集中する性質を示す．これを用い，低周波成分のパワーを閾値処理することにより，雑音下でも 90-95%の精度で VAD が可能である．ただし，マイクの装着法，発話者の声量によって性能はばらつく．

---

*Introduction of speech technology in service-operation estimation.
by TAKEHARA, Masanori, TAMURA, Satoshi (Gifu Univ.), TENMOKU, Ryuhei, KURATA, Takeshi (AIST), HAYAMIZU, Satoru (Gifu Univ.)

発話区間情報は位置情報や SO と統合して有用な統計量となる．例えば，客室や客席の前での発話は顧客との会話として判別され，顧客とどれだけコミュニケーションをとったかの指標となる．

3.2 キーワードスポッティング

従業員の発話内容は従業員の行動や状況と相関があると考えられる．例えば,「ご注文はお決まりですか.」と従業員が発話した場合，発話者がこれからオーダーをとろうとしている状態，顧客が注文を頼もうとしている状態を表す．このような接客の決まり文句（以下，キーワード）を正しく認識できれば，様々な行動や状況を推定できることになる．ただし，推定の必要な 1 つの動作に対して発話内容や言い回しは複数存在する．このような場合，大語彙音声認識よりも発話中のキーワードに焦点を当てた認識の方が有用と考えられる．先ほど挙げた例では「ご注文」や「お決まり」がキーワードとなる．これらのキーワードは SO のパラメータの 1 つとなる他，可視化によって吹出し型でリアルタイムに表示することで業務のフィードバック支援が期待される．

音声データには雑音や他者の発話の重畳が随所に見られるため，認識は容易ではない．音素認識の結果からキーワードの探索を行う，キーワードの音響モデルとガベージの音響モデルを作成しキーワードの認識を行うといった方法を検討している．

4 SOE 実験

4.1 実験方法

筆者らは，日本食レストラン，高齢者介護施設等のサービス現場にて従業員の行動計測を行った．各種センサ，骨導音マイクを 1 日約 10 人の従業員に装着してもらい，10 時間前後連続してデータを収集する．計測は 2ヶ月間行われ，合計 5000 時間の測位データ，音声データを収集した．

SOE は要素データとして 37 次元の特徴量 [2] を用意し Adaboost [3] によって学習を行い分類器を構築した．そして，8 種類の SO（注文伺い，配膳，会計など）の識別を行った．前節で推定した発話区間情報は発話量・非発話量の算出に用いる．発話量は式 (1) で示すように一定区間内の発話区間（または非発話区間）の割合を指し．それぞれ閾値以上であるか否か（0 または 1）を 2 次元の特徴量として利用している．なお，式 (1) の分析窓長は 25[s] である．

$$\text{発話量} = \frac{\text{発話と判定された時間長 } [s]}{\text{分析窓長 } [s]} \qquad (1)$$

4.2 実験結果

Table 1 は各 SO に対する弱識別器（発話量・非発話量）の寄与率と，37 個全ての弱識別器中の寄与率の順位である．発話量以外の主な弱識別器として，鉛直・水平動作の有無，各エリアの滞在時間，エリア通過数等がある．移動・運搬という発話行為の少ない

Table 1 Contribution of speech and non-speech rates in SOE.

	発話量		非発話量	
	寄与率	順位	寄与率	順位
注文伺い	11.6%	3	9.1%	5
配膳	11.3%	1	3.4%	11
移動・運搬	0.1%	26	4.2%	10
会計	9.8%	3	8.4%	5
挨拶・案内	8.7%	4	13.0%	1
片付け・準備	4.4%	12	8.7%	2
お客さんと会話	21.2%	1	9.3%	3
スタッフと会話	8.4%	4	0.0%	20

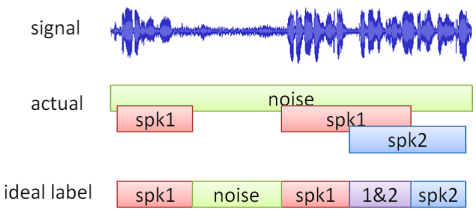

Fig. 2 A VAD example for multiple speekers and noises.

SO を除いた 7 種類の SO で，発話量が高い寄与率を示しており発話量の有効性が確認出来る．

このように，音声データから発話区間を求めるだけでも行動の識別に有用である．実データはマイク装着者の発話だけでなく，Fig.2 に示すように他者の発話の重畳も発生する．VAD 中に話者識別を行うことができれば，SOE の精度向上にさらに有効であると期待される．

5 まとめ

本稿は，SOE に VAD とキーワードスポッティングの技術を導入する試みと，その可視化について紹介した．また，骨導音データの VAD によって算出した発話量が SOE に有用な特徴量であることを示した．今後は，各々の推定アルゴリズムを改良し要素データの精度を上げること，複数の話者を識別する VAD の枠組みを構築し SOE に有用な要素データを新たに生成することを課題として取り組んで行く．

参考文献

[1] 天目隆平，竹原正矩，他：労働集約型サービス—従業員行動計測技術に基づく分析と可視化，HCG シンポジウム 2010 論文集，pp.443-448，2010.

[2] 天目隆平，上岡玲子，他：日本食レストラン産業におけるマルチセンサと POS データに基づくサービスオペレーション推定，FIT2011 論文集, 2011.

[3] Y. Freund and R. E. Schapire: ”A Decision-theoretic Generalization of On-line Learning and an Application to Boosting,” Jour. of Computer and System Sciences, Vol. 55, No. 1, pp. 119-139, 1997.