話速の変動を捉える特徴量に基づく留守録音声の緊急度推定*

◎神山歩相名，安藤厚志，増村亮，小橋川哲，青野裕司
(日本電信電話株式会社 NTT メディアインテリジェンス研究所)

1　はじめに

コンタクトセンタ（例えば，故障受付等）の中には，夜間等のコスト削減のために顧客の要望を録音音声（留守録）で一次受けを行い，順に対応することが行われている．このような録音受にといて顧客にとっては，緊急性が高いときは，優先的に対応してというニーズある．緊急性が高いとき顧客は，発話の仕方（音声大きさ，高さ，速さ等）が非緊急の時と比べて異なる．そこで，本研究では音声の発話の仕方から，緊急度の判定（緊急度 高, 低の 2 クラス識別）するタスクに取り組む．

2　従来手法

本研究では，時系列特徴量を入力するニューラルネットワークを用いて緊急度を推定する．

2.1　問題設定

本研究では，音声から留守録音声の緊急性を判定する．判定クラスは,「緊急」「非緊急」の 2 クラスである.「緊急」は，他の留守録より優先的に返信する必要がある留守録,「非緊急」は，優先に対応しなくても良い留守録を示す．留守録の「緊急」「非緊急」の推定は，次のように定式化ができる．

$$\hat{c} = \arg\max_{c \in C} p(c|\mathbf{U}, \mathbf{\Theta}) \tag{1}$$

$\hat{c}$ は推定された緊急度，C はクラスラベルの集合を表し $C \in 0, 1$ とする（$c = 0$ のとき「非緊急」，$c = 1$ のとき「緊急」を示す）．$p(c|\mathbf{U}, \mathbf{\Theta})$ は緊急度推定モデルを示し，$\mathbf{U}$ は緊急度推定のための特徴量セットを示し，$\mathbf{\Theta}$ はパラメータセットを示す．

2.2　緊急度推定モデル

従来は，Mel-frequency Cepstral Coefficients (MFCC)，基本周波数 (F_0) 等の音響特徴量の統計量と，音声全体の平均話速を用いて緊急度を推定している [1, 2]．従来は音響特徴量の統計量と話速のベクトルを Support Vector Machine 等で緊急度を推定していたが，近年は感情認識等で，数十ミリ秒単位で抽出した音響特徴量系列を直接入力するニューラルネットワークが提案されている [3]．緊急度推定において，ミリ秒単位で抽出した音響特徴量系列（以下,「短時間特徴量」とする）と音声全体の話速情報（以下,「全体平均特徴量」ニューラルネットワークを Fig. 1 に示す．モデルは，次のような式で表現することができる．

$$\mathbf{h}_X = \mathtt{RNN}(\mathbf{x}_1, \cdots, \mathbf{x}_{T_1}; \mathbf{\Theta}_X) \tag{2}$$

$$p(c|\mathbf{U}, \mathbf{\Theta}) = \mathtt{SOFTMAX}(\mathbf{w}_1[\mathbf{h}_X, y] + \mathbf{b}_1) \tag{3}$$

特徴量セットは $\mathbf{U} = (\mathbf{x}_1, \cdots, \mathbf{x}_{T_1}, y)$ であり，短時間特徴量 $\mathbf{x}_1, \cdots, \mathbf{x}_{T_1}$ および全体平均特徴量 y を持つ．$\mathtt{RNN}(\cdot)$ は Recurrent Neural Network (RNN) であり，RNN のモデルパラメータ $\mathbf{\Theta}_X$ を持つ．RNN は，短時間特徴量 $\mathbf{x}_1, \cdots, \mathbf{x}_{T_1}$ を入力し，短時間特徴量の情報を埋め込んだ隠れベクトル $\mathbf{h}_X$ を出力する．y は全体平均特徴量，$[\mathbf{h}_X, y]$ はベクトル $\mathbf{h}_X$ と全体平均特徴量 y の結合ベクトルである．$\mathbf{w}_1$,$\mathbf{b}_1$ はモデルパラメータであり，$\mathtt{SOFTMAX}(\cdot)$ は softmax 関数である．これにより，緊急度の高い音声の声の高さ，大きさ等の局所的な変化と，全体の平均話速の違いの特徴を用いて緊急度を推定することができる．

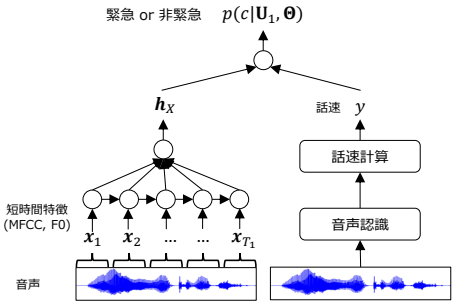

Fig. 1　緊急度推定モデル

3　提案手法

本節では，緊急度の音声のリズムの変化を用いた手法について説明する．

3.1　アプローチ

我々は，緊急と非緊急の音声は話者の冷静さが異なり，緊急時には感情的に話すため，話速の変動（以下,「リズム」とする）が大きいと考えた．従来手法は，ミリ秒単位の短時間特徴量であるため，リズムを捉える単位としては短すぎて，モデルにリズムの違いを学習させるのは困難である．一方，話速のような全体平均特徴量では，リズムまで扱うことができない．

そこで本研究ではリズムの特徴を捉えるために，長時間の窓幅を用いた特徴量抽出を行い，緊急度推定を行う．提案手法は，全体平均特徴量のかわりに，秒単位の窓幅のリズムに関する特徴量（以下,「長時間リズム特徴量」とする）を抽出し，短時間特徴量と組み合わせて緊急度を推定する．リズムに関する長時間リズム特徴量として，振幅のゆるやかな変動である Envelope Modulation Spectrum (EMS) および音の変動の大きさを示す MFCC の時間ごとの統計量（平均，標準偏差，歪度，尖度，標準絶対偏差，最大値，以下「MFCC-stat」とする）を用いる [4]. EMS はオクターブバンドフィルタ毎の音声信号の緩やかな振幅の変動を求め，秒単位の窓幅で各変動をパワースペクトルに変換し，そのパワースペクトルの最大パワー値とその周波数，帯域間のパワー比を特徴量とする．この秒単位のスケールの振幅変動を特徴とするためリズムに関する特徴とすることができ，構音障害等の検出に用いられている [5].

3.2　提案手法

提案手法は，短時間特徴量と長時間リズム特徴量の抽出を行い，ニューラルネットワーク上で結合し，緊急度を推定する．このモデル構造を Fig. 2 に示す．

---

*Urgency voicemail detection based on variability of speech rate by Hosana KAMIYAMA, Atsushi ANDO, Ryo MASUMURA, Satoshi KOBASHIKAWA and Yushi AONO (NTT Media Intelligence Laboratories, NTT Corporation)

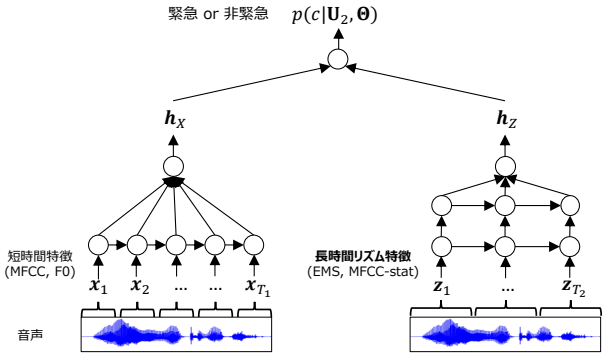

Fig. 2　提案するニューラルネットワーク

モデルは次のようになる．

$$\mathbf{h}_X = \texttt{RNN}(\mathbf{x}_1,\cdots,\mathbf{x}_{T_1};\mathbf{\Theta}_X) \quad (4)$$

$$\mathbf{h}_Z = \texttt{RNN}(\mathbf{z}_1,\cdots,\mathbf{z}_{T_2};\mathbf{\Theta}_Z) \quad (5)$$

$$p(c|\mathbf{U},\mathbf{\Theta}) = \texttt{SOFTMAX}(\mathbf{w}_2[\mathbf{h}_X,\mathbf{h}_Z]+\mathbf{b}_2) \quad (6)$$

特徴量セットは $\mathbf{U}=(\mathbf{x}_1,\cdots,\mathbf{x}_{T_1},\mathbf{z}_1,\cdots,\mathbf{z}_{T_2})$ であり，短時間特徴量 $\mathbf{x}_1,\cdots,\mathbf{x}_{T_1}$，および長時間リズム特徴量 $\mathbf{z}_1,\cdots,\mathbf{z}_{T_2}$ を持つ．$\mathbf{\Theta}_Z$ は，長時間リズム特徴量を入力する RNN のパラメータであり，$\mathbf{h}_Z$ は，長時間リズム特徴量系列を埋め込んだ隠れベクトルである．$[\mathbf{h}_X,\mathbf{h}_Z]$ は，ベクトル $\mathbf{h}_X$ およびベクトル $\mathbf{h}_Z$ を結合したベクトル，$\mathbf{w}_2$,$\mathbf{b}_2$ はモデルパラメータである．短時間特徴量と長時間リズム特徴量は窓幅が異なるため，RNN を用いて得られた 2 つのベクトルを結合し，緊急度を推定する．

4　実験

提案手法の有効性を確認するため，評価実験を行った．

4.1　データセット

本稿では，冷凍食品販売の夜間受付を想定して，留守録音声を収録した．初めに緊急・非緊急の状況設定・留守録に入力する要件を定めたシナリオを作成し，各話者シナリオに沿った内容で自由に発話を行って収録を行った．話者は 20 名でそれぞれ 12 シナリオ収録した．各音声の平均時間長は約 30 秒で，8kHz サンプリングで収録した．

続いて，各音声について 3 名のアノテータにて緊急・非緊急のラベルを付与し，3 名とも緊急・非緊急のラベルが一致したデータのみ実験に用いた．3 名のラベルの一致率である Cohen の κ 係数は 0.89 で，緊急音声 120 サンプル，非緊急音声 100 サンプルが得られた．

4.2　実験条件

本研究では，短時間特徴量，長時間リズム特徴量および全体平均特徴量の 3 つの特徴量の比較を行った．短時間特徴量は，パワー項を含む 13 次元の MFCC および F_0 を抽出し，それらの 1 次微分を含めて 28 次元の特徴量を抽出した．抽出条件は，窓幅 25 ミリ秒，シフト幅 10 ミリ秒で抽出した．長時間リズム特徴量は，リズムに関係する特徴量として EMS，1 秒間の MFCC-stat を抽出した [4]．EMS の抽出条件を，Table 1 に示す．EMS と MFCC-stat は，共に窓幅 1 秒，シフト幅 100 ミリ秒シフトで抽出した．全体平均特徴は，音声認識結果（ASR）および書き起こしテキスト（Oracle）を用いて話速を算出して，緊急度推定に用いた．

短時間特徴量の RNN には，32 次元のベクトルを出力する注意機構付き Long-Short Term Memory (LSTM) を用いた．提案手法の，長時間リズム特徴量の RNN には，64 次元のベクトルを出力する 2 層の注意機構付き LSTM を用いた．最適化関数は，Adam を用いて学習率は 0.001 を用いた．

緊急度推定は 10 分割の Cross-validation を行い，9/10 を学習データ，1/10 を評価データとした．各学習セットで 5 回評価を行い，最も高い Accuracy で比較を行った．

Table 1　EMS 抽出条件

オクターブバンド中心周波数 [Hz]	30, 60, 120, 240, 480, 1920, 3480
特徴量	最大パワーの周波数 最大パワー 3-6 Hz 帯域パワー 0-4 Hz 帯域パワー 4-10 Hz 帯域パワー 0-4 Hz と 4-10 Hz 帯域のパワー比

Table 2　緊急度推定結果

短時間	長時間リズム		全体平均	Acc.
MFCC,F_0	EMS	MFCC-stat	話速	
✓	-	-	-	.748
✓	-	-	✓(ASR)	.776
✓	-	-	✓(Oracle)	.790
-	✓	-	-	.709
-	-	✓	-	.732
-	✓	✓	-	.755
✓	✓	-	-	.791
✓	-	✓	-	.886
✓	✓	✓	-	.895

4.3　結果

各特徴量における評価結果を Table 2 に示す．提案する短時間特徴量と，長時間リズム特徴量を用いた結果が 89.5%と最も高い推定精度となった．従来の短時間特徴量と全体特徴量話速を用いた結果と比べて 50%の誤り削減率を達成した．長時間リズム特徴量単体では，短時間特徴量を用いた時よりは低い推定精度となった．これは，緊急の音声の急峻な声の高さ，大きさ，声質の変化も緊急度推定には重要であることを示している．

5　まとめ

本稿では, 緊急時の音声のリズムの変化に着目して，留守録音声の緊急度推定を行った．従来は，数十ミリ秒単位の短時間特徴量，および通話全体の平均話速である全体平均特徴量を緊急度推定に用いていたが，緊急時のリズムの変動のモデル化ができなかった．提案手法は，リズムに関連する秒単位の長時間リズム特徴量を抽出し，ニューラルネットワーク上で結合して緊急度を推定した．実験の結果，リズムに関連する長時間リズム特徴量の有効性を確認した．今後は，テキスト情報も用いた緊急度推定について検討する必要がある．

参考文献

[1] Z. Inanoglu *et al.*, *Proc. IUI*, 2005.
[2] 堀 他, 信学技報, SP2017-95, 2018.
[3] S. Mirsamadi *et al.*, *Proc. ICASSP*, 2017.
[4] Y. Jiao *et al.*, *Proc. ICASSP*, 2016.
[5] J. Liss, *et al.*, *Journal of Speech Language and Hearing Reserch*, 2010.