ISS-SP-211

状況による話者内音声変動に頑健な話者識別手法

橋本　哲平† 　堀内　靖雄† 　黒岩　眞吾†
† 千葉大学　大学院融合科学研究科

1. はじめに

話者識別において，話者の感情や気分が変化することで登録時と識別時の音声に音響的な違いが生じる話者内音声変動によって，識別性能が劣化する問題がある[1]．この問題を解決するため，本稿ではGMM-SVMによる話者識別システム[2]を対象として，推定した話者内音声変動を学習データに付加することで，学習データを拡張し識別性能を向上させる手法を提案する．

2. 予備実験

ボードゲーム音声を用いて感情や気分による話者内音声変動が話者識別システムへ与える影響を調査する．この音声は21名のゲーム中の指示文の読み上げ音声248発話と，ゲーム前後のATR音素バランス文Aセット10文(A01~A10)の読み上げ音声を収録している．聴取の結果，ゲーム開始前の読み上げ音声と比較して，ゲーム中やゲーム後では音声に変動が生じていることを確認した．このうち11名の話者の各場面の5発話を登録，残りの発話を評価に用いて以下の2つを調査した．

(1) 登録音声にゲーム前の音声（A01~A05)を使用し，評価音声をゲーム前（A06~10)，ゲーム中（A06~10)，ゲーム後と変化させたときの識別性能の変化

(2) 登録音声の違い(ゲーム前(A01~A05)，ゲーム前+中，ゲーム前+後(A01~A05)，ゲーム前+中+後)による，ゲーム中の音声の識別性能の変化

調査結果を表1，2に示す．表1よりゲーム中の音声を評価した場合の識別性能が最も低く，ゲーム中の音声に最も大きな変動が生じていると考えらえる．また表2より，話者モデルの学習にゲーム中やゲーム後の音声を加えることで頑健な話者モデルを得られることが分かった．以上のことからゲーム前の学習データに対して同様の変動を学習データに付加することで頑健な話者モデルを得られると考えた．

3. 提案手法

提案手法では，予備実験に用いていない残りの話者(参照話者)10名のゲーム前とゲーム中，ゲーム前とゲーム後のGMM-SuperVector(SV)間の差分を変動としてGMM尤度を用いた重み付け和によって登録者のGMM-SVに付加することで，擬似的に登録者のゲーム中やゲーム後の学習データを得る．

ゲーム後の学習データは次の式によって求めた．

$$\Delta'^{r}_{k} = \boldsymbol{\varphi}'^{r}_{k} - \boldsymbol{\varphi}^{r}_{k} \tag{1}$$

表1　(1)話者内音声変動による影響

学習条件	評価音声	誤識別率
ゲーム前	ゲーム前	0%
	ゲーム中	11.2%
	ゲーム後	4%

表2　(2)変動に対する頑健性

学習条件	評価音声	誤識別率
ゲーム前	ゲーム中	11.2%
ゲーム前+後		4.9%
ゲーム前+中		2.1%
ゲーム前+中+後		0.0%

表3　実験結果

学習条件	評価音声	誤識別率
提案手法	ゲーム中	8.4%

$$\boldsymbol{\varphi}'^{s}_{k} = \boldsymbol{\varphi}^{s}_{k} + \frac{\sum_{r\epsilon R} p(\boldsymbol{X}^{s}_{k}|\boldsymbol{\lambda}^{r}_{k})\,\Delta'^{r}_{k}}{\sum_{r\epsilon R} p(\boldsymbol{X}^{s}_{k}|\boldsymbol{\lambda}^{r}_{k})} \tag{2}$$

またゲーム中の学習データは次の式によって求めた．

$$\Delta''^{r}_{all} = \boldsymbol{\varphi}''^{r}_{all} - \boldsymbol{\varphi}^{r}_{all} \tag{3}$$

$$\Delta^{r}_{k-all} = \boldsymbol{\varphi}^{r}_{k} - \boldsymbol{\varphi}^{r}_{all} \tag{4}$$

$$\Delta'^{r}_{k-all} = \boldsymbol{\varphi}'^{r}_{k} - \boldsymbol{\varphi}'^{r}_{all} \tag{5}$$

$$\Delta''^{r}_{k} = (\Delta''^{r}_{all} + \Delta^{r}_{k-all}) \text{ or } (\Delta''^{r}_{all} + \Delta'^{r}_{k-all}) \tag{6}$$

$$\boldsymbol{\varphi}''^{s}_{k} = \boldsymbol{\varphi}^{s}_{all} + \frac{\sum_{r\epsilon R} p(\boldsymbol{X}^{s}_{k}|\lambda^{r}_{k})\,\Delta''^{r}_{k}}{\sum_{r\epsilon R} p(\boldsymbol{X}^{s}_{k}|\lambda^{r}_{k})} \tag{7}$$

ここで，sは登録者，rは参照話者，Rは参照話者rの集合を表す．また，$\boldsymbol{\varphi}$，$\boldsymbol{\varphi}'$，$\boldsymbol{\varphi}''$はそれぞれゲーム前，後，中のGMM-SV，$\boldsymbol{\varphi}_{all}$は全学習発話から抽出したGMM-SV，$\boldsymbol{\varphi}_{k}$は発話内容$k$の学習発話から抽出したGMM-SV($\boldsymbol{\varphi}_{all}$も含む)，$p(\boldsymbol{X}^{s}_{k}|\lambda^{r}_{k})$は$s$の学習発話$k$の特徴ベクトル列$\boldsymbol{X}^{s}_{k}$と$r$の学習発話$k$から学習したGMM λ^{r}_{k}との尤度を表す．

4. 話者識別実験

実験結果を表3に示す．提案手法では，表2のゲーム前の音声で学習した場合と比較して誤識別率を2.8ポイント改善した．

5. 今後の課題

今後，より効果的な変動ベクトルの抽出と付加方法を検討する必要がある．

謝辞：本研究の一部は，JSPS科研費JP16K00229及び千葉大学VBL研究プロジェクトの助成を受けたものです．

参考文献

[1] M.V.Ghiurcau et al, SPAMEC 2011 pp.81-84Aug 2011
[2] D.A.Reynolds et al, IEEE Signal Processing Letters, vol. 13, no. 5, pp308-311, May 2006.