Vol.2018-MUS-119 No.50
Vol.2018-SLP-122 No.50
2018/6/17

正規化特徴量を用いた多言語対応感情識別法の提案

岡部 成美[1]　川村 将士[1]　鈴木 基之[1,a)]

概要：本研究では正規化特徴量を用いて音声から多数の言語に対応した感情識別を行う方法を提案する．一般に感情識別器は対象言語をひとつに設定している．これは言語が異なるとアクセントやイントネーション等の違いにより同じ感情でも特徴量が変化してしまうからである．そこで本研究では，感情を含む音声の特徴量と感情を含まない音声（合成音声）の特徴量との差分をとることで正規化を行い，言語によらず感情による変異のみを抽出することで，ひとつの識別器で多数の言語の音声から感情を識別する方法を提案する，これを用いて対象言語を 4 言語にして識別器の性能評価を行った．対象言語のみで正規化なしの場合，識別率は平均して 60.6 %，一方 4 言語で正規化ありの場合は平均して 58.9 %と両者にあまり差はなく，既知言語の場合は多言語化した識別器を用いてもほとんど性能劣化がないことが分かった．一方，学習言語に含まれない言語では性能が出ないことから，完全には言語の違いを吸収できないと言える．

1. はじめに

近年，情報社会に伴うコンピュータの発達によりスマートフォン等の小型端末が広く普及している．その中で音声対話システムが多く開発されており，例として Siri[1] などのアプリケーションや Google Home[2] などの AI スピーカーなどが挙げられ，すでに実用化されている．これらのシステムには音声認識や音声合成の技術が用いられており，実用できる程度に技術は発達しているが，感情音声識別や感情音声合成の技術はまだ実用化には至っていない．

電話等で会話を行う際に人間は言葉の内容だけでなく喋り方によって感情を汲み取とり，相手の感情によって話をさらに展開したり内容を変更したりということを自然に行っている．このように感情は円滑なコミュニケーションを行うための重要なツールであると言える．しかし先ほど挙げたような音声対話システムには音声から感情を判断するという機能は備わっていない．例えば「遊びに行くのにお勧めの場所はある？」と発話したときにシステム側は「動物園はどうですか？」と返答したとする．それに対して「動物園かぁ」と発話するとシステム側は発話者が動物園に行きたいかどうか判断できない．もし感情音声識別ができると喜びと判断した場合には「ここの動物園が人気です」とさらに詳しい情報を提示したり嫌悪を示している場合は「水族館はどうですか？」と別の提案をしたりすることができる．このように自然な対話が可能となるため，感情音声識別技術の発展が期待されている．

現在，感情識別の研究は数多く行われている（例えば [3][4] 等）が，これらは 1 つの言語を対象としている．これは対応した言語以外で感情識別を行おうとすると識別率が著しく低下してしまうからである．例えばドイツ語で学習した識別器でドイツ語の感情識別を行うと識別率は 73.9%であるが，この識別器で日本語の感情識別を行うと 20.5%となり識別率は 53.4 ポイントも低下してしまう．これは，言語によってアクセントやイントネーション等が異なり，同じ感情でも特徴量が変化してしまい，上手く認識が行えないからである．しかし仮に識別器の多言語化が可能になると言語によって識別器を変える必要がなくなり，また未知の言語においても既知の言語と同等の感情識別が可能になると予想される．そこで，本論分では中川 [5] が提案した発話内容に依存しない感情識別法を応用して言語に依存しない感情識別機の構築を目指す．中川は喜びや怒りなどの感情は平静の感情からどれだけ韻律が変化したかという部分に表れると考え，合成音声を平静の感情による同内容の発話として用い，感情を含む音声との差分を特徴量として用いることでより高精度に感情識別が可能になることを示した．同じように言語によってもアクセントやイントネーションが異なるため，先ほどと同じように正規化を用いることよって言語による影響も吸収できると思われる．本論文ではこの方法を用いることで言語に依存しない感情識別器を構築してその性能を評価する．

[1] 大阪工業大学情報科学部
Osaka Institute of Technology
[a)] moto@m.ieice.org

Vol.2018-MUS-119 No.50
Vol.2018-SLP-122 No.50
2018/6/17

2. 正規化特徴量を用いた多言語対応感情識別法

2.1 発話内容に依存しない感情識別法

中川が提案した方法 [5] を図 1 に示す．一般的には感情を含む音声から特徴量系列を出力し，それらの統計量からベクトルを作成してそれらを学習させて識別を行う．しかし中川の方法では感情音声と合成音声の特徴量の差分をとり，得られた系列に対して統計量を計算するということを行っている．具体的には感情音声が入力されるとその音声フレームごとに分割して，各フレームで特徴量を抽出する．さらに同じ発話内容である平静感情の音声に対しても同様に特徴量を抽出する．その後両者の特徴量を用いて各フレームごとに正規化を行い，発話文全体で最大値や分散値等の統計量を計算した結果を識別器に学習させる．正規化で用いる平静音声であるが，感情音声を発話したときに同じ発話者の同じ発話内容の平静音声を同時に収録することは不可能であるため，合成音声を用いることとする．同じ発話内容の合成音声を生成するにはテキストを用意する必要があるため，その際に音声認識を行う．この方法を用いることで発話内容に依存しない感情識別が可能となった．

2.2 時間波形正規化法

正規化は感情音声と平静音声の各フレームごとの特徴量を用いて行うが，双方の発話時間が違い，フレーム数が異なってしまうためフレームの対応付けを行う必要がある．そこで DP マッチングを用いる．このときに発話音声から抽出された特徴量の 1 つである「MFCC」を使用する．この特徴量は声道情報を表した特徴量であり，「何を喋っているか」ということを示すものであり，対応付けを行うのに有効である．DP マッチングの概要を図 2 に示す．図 2 の Frame(E) は発話した感情を含む音声であり，Frame(N) は合成器で作成した感情を含まない音声である．例えばこの図 2 ではフレームの対応付けを行った結果，感情音声のフレームの 1 個目は平静音声の 1 個目と 2 個目が対応するというようになっている．

正規化は減法または除法を特徴量の種類ごとに適切に選択して行う．例えば「基本周波数」は倍になると 1 オクターブ高くなり，半分になると 1 オクターブ低くなることから，正規化によって感情音声と平静音声の差を求める際には減法より除法を用いた方が良いと考えられる．このように特徴量によって最適な方法が異なると考えられるため，「正規化なし」「除法で正規化」「減法で正規化」のうちから最適な正規化を行うようにした．減法と除法についての計算式は以下の通りである．

$$f_s(x_i) = x_i - \frac{\sum_{y_i \in c(x_i)} y_i}{|c(x_i)|} \quad (1)$$

$$f_d(x_i) = \frac{1}{|c(x_i)|} \cdot \sum_{y_i \in c(x_i)} \frac{x_i}{y_i} \ (|y_i| >= d) \quad (2)$$

$$f_d(x_i) = \frac{1}{|c(x_i)|} \cdot \sum_{y_i \in c(x_i)} \frac{x_i}{d} \cdot sign(y_i) \ (|y_i| < d) \quad (3)$$

$$f_d(x_i) = \frac{1}{|c(x_i)|} \cdot \sum_{y_i \in c(x_i)} \frac{x_i}{d} \ (|y_i| = 0) \quad (4)$$

式 (1) は減法の際に用いて式 (2)〜(4) は除法の際に用いる．減法の場合は感情音声の特徴量を平静音声の特徴量で減算する．対応するフレームが複数存在する場合はそのフレーム分の平均をとった値を使用する．除法の場合は対応付けしたフレームごとにそれぞれ除法を行い，そのフレーム分の平均を計算する．x_i は感情音声のフレームごとの特徴量であり，y_i は平静音声のフレームごとの特徴量である．$c(x_i)$ は対応付けしたフレームの集合であり，例えば図 2 で説明すると集合 E(1) には N(1) と N(2) の要素が含まれているということになる．$|c(x_i)|$ は $c(x_i)$ に含まれるフレーム数であり，同じく図 2 で説明すると集合 E(1) には 2 個の要素が含まれていることになる．d は十分小さい正の値である．また符合を考慮するために $sign(y_i)$ とすることで符号をとっている．

2.3 言語に依存しない感情識別法の提案

今回提案する方法を図 3 に示す．まず平静音声を作成するために言語識別を行い，言語ごとに音声認識や音声合成を行う必要がある．平静音声が作成できた後は中川と同様に特徴量を計算する．そして全ての言語の特徴量を合わせて，1 つの識別器を学習する．

3. 性能評価実験

3.1 使用コーパス

本実験では日本語，ドイツ語，英語，イタリア語の 4 言語を用いた．各コーパスの概要を表 1 に示す．

日本語のコーパスは感情評定値付きオンラインゲーム音声チャットコーパス [6] を使用する．男女 2 人ずつの計 4 人が対話形式で演技した音声が収録されている．今回は他のコーパスと感情の種類を統一するために「喜び」「嫌悪」「恐れ」「悲しみ」「怒り」の 5 つの感情のみを使用するので全部で 1224 発話となる．1 発話に対して感情の強度は感情のない音声を強度 0，弱を強度 1，中を強度 2，強を強度 3 として感情が 4 段階に設定されている．1 人につき 102 文発話しており，それぞれ強度を変えて発話しているため 1 人の発話数は 306 発話となる．感情によって発話内容はすべて異なるが発話者によって発話内容が異なることはない．

ドイツ語は Berlin Database of Emotion Speech[7] を使用する．男女 5 人ずつの計 10 人が演技した音声が収録さ

Vol.2018-MUS-119 No.50
Vol.2018-SLP-122 No.50
2018/6/17

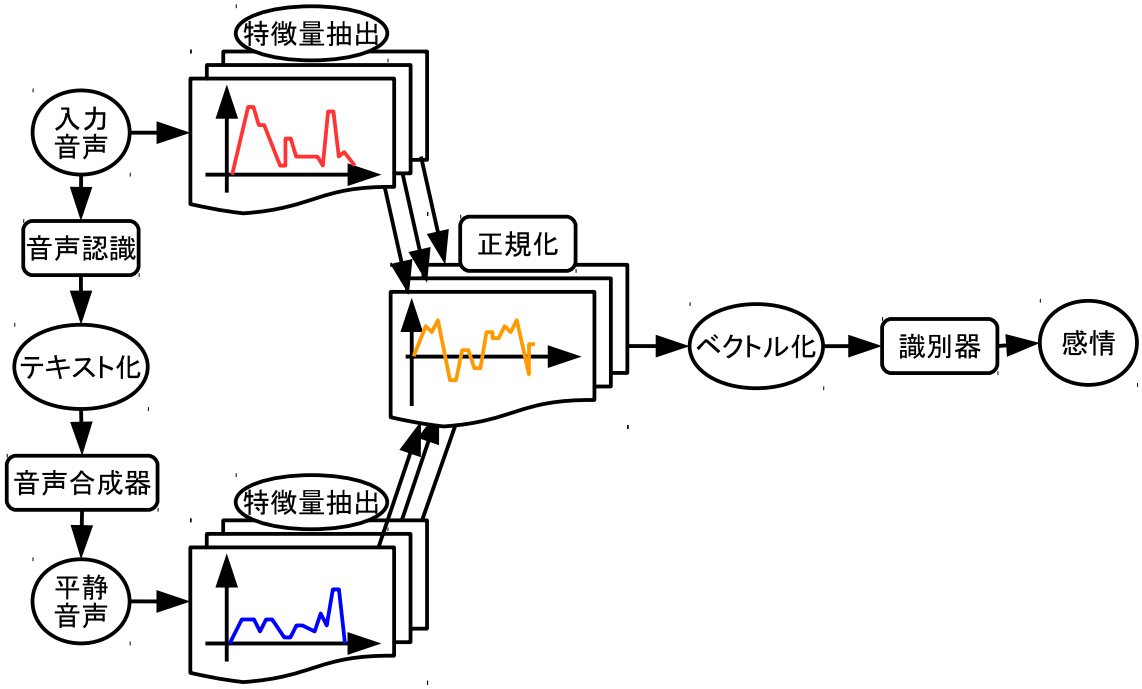

図 1　正規化を用いた感情識別法

表 1　コーパスの内訳

データベース		OGVC	Emo-DB	SAVEE	EMOVO
話者数		男性 2 名女性 2 名	男性 5 名女性 5 名	男性 4 名	男性 3 名女性 3 名
発話数		1224 発話	375 発話	300 発話	420 発話
音声の種類		演技音声	演技音声	演技音声	演技音声
感情強度		1(弱)-2(中)-3(強)	なし	なし	なし
発話文数	喜び	252 発話	70 発話	60 発話	84 発話
	嫌悪	240 発話	47 発話	60 発話	84 発話
	恐れ	240 発話	70 発話	60 発話	84 発話
	怒り	240 発話	127 発話	60 発話	84 発話
	悲しみ	240 発話	61 発話	60 発話	84 発話

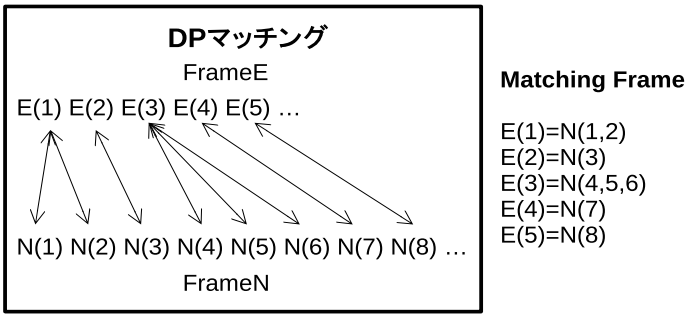

図 2　フレームの対応付け

れている．今回は他のコーパスと感情の種類を統一するために先ほどと同じ 5 感情を用いるので使用するのは全部で 375 発話となる．発話者により 1 発話の中に含まれている感情の数が違うため，1 人あたりの発話数が異なる．また感情にまたがって同じ内容の発話を行っている場合もあり，それも発話者によって異なる．

英語は Surrey Audio-Visual Expressed Emotion Dtabase[8] を使用する．男性 4 人が演技した音声が収録されている．今回は他のコーパスと感情の種類を統一するために先ほどと同じ 5 感情を用いるので使用するのは全部で 300 発話となる．発話者ごとの発話数は同じであり，1 人 1 感情において 15 文ずつである．この 15 文の発話内容はすべて異なり，感情ごとにもほぼ異なっているが 3 文のみが全感情で共通している．また感情ごとに発話文は異なるが 3 文だけはどの感情でも同じ発話内容のものが使用されている．

イタリア語は EMOVO Corpus[9] を使用する．男女 3 人ずつの計 6 人が演技した音声が収録されている今回は他のコーパスと感情の種類を統一するために先ほどと同じ 5 感情を用いるので使用するのは全部で 420 発話となる．. 発話者ごとの発話数は同じであり，1 人 1 感情において 14 文ずつである．この 14 文の発話内容はすべて異なるが全感情で共通している．また感情をまたいでも同じ内容の文が発話されている．

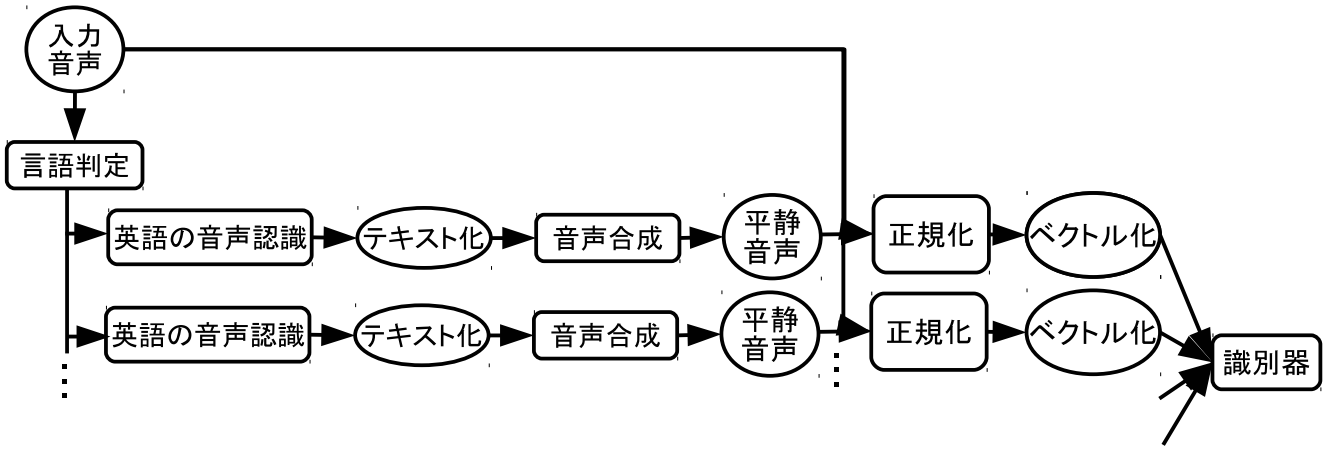

図 3 多言語対応感情識別法

3.2 平静音声の作成

平静音声は音声合成で生成することで簡単に用意できる．今回使用した音声合成器は Open JTalk[10] と Text To Speech[11] である．Open JTalk は日本語の平静音声を作成する際に使用して，性別は感情音声に合わせて同じものを用いた．Text To Speech はドイツ語，英語，イタリア語の平静音声を作成する際に使用して，声の性別は特に注視せず今回は男性で統一した．また合成音声を作成するために，言語識別と音声認識を行う必要があるが，認識誤りが発生して正規化を行う場合の正しい効果を検証することができないため，本論文では音声認識を行わず人間が手で書き起こした発話文を合成器に渡すことで合成音声を作成した．

3.3 韻律的特徴量

本実験では The INTERSPEECH 2009 Emotion Challenge[12] で提案された，感情識別を行う際に妥当であるとされる韻律的特徴量の「平均パワー」「ゼロクロス比」「VoiceQuality」「基本周波数」「MFCC(12 次元)」の 5 つの特徴量を用いた．それぞれフレームごとに抽出した静時特徴量とその変位を表した動的特徴量を用いて特徴量ごとに「正規化なし」「除法で正規化」「減法で正規化」の 3 パターンで正規化を行い，それぞれ 12 種類の統計量（最大値，最小値，範囲，最大値のフレーム番号，最小値のフレーム番号，平均値，線形回帰式の傾き，線形回帰式の切片，分散値，歪度，尖度）をとり，最終的に 384 次元（$=(16+16)\times 12$）のベクトルに変換した．正規化の方法は静的特徴量と動的特徴量のそれぞれで 5 つの特徴量の正規化パターンを全て検証して，特徴量によってどの方法を用いるのが適切かを検証して最も最適な方法を用いるべきであるが，全パターンを検証しようとすると $3^{10}=59049$ 通りとなるため，現実的ではない．そこで今回は静的特徴量と動的特徴量の正規化パターンを固定して正規化を行った．

表 2 各言語の認識率

	正規化なし	正規化あり
日本語	48.6 %	49.5 %
ドイツ語	73.9 %	74.9 %
英語	34.7 %	41.3 %
イタリア語	85.2 %	86.7 %

3.4 学習・評価方法

本実験では識別器として libSVM[13] で実装されている SVM(Support Vector Machine) を用いた．SVM の学習と評価であるが，本実験で用いるコーパスはデータ数が少ないため，クロスバリデーションを用いた．日本語，ドイツ語，英語は 20 グループ，イタリア語は 24 グループに分割した．またグループによって感情ごとの発話数がなるべく同じになるようにグループ分けを行った．

SVM のパラメータは全データを用いたグリッドサーチによって決定し，その値を用いてクロスバリデーションで結果を求めた．

3.5 言語独立の実験と結果

まずは各言語の特性を調べるために，言語ごとに識別器を作成して対応した言語で感情識別を行った．その結果を表 2 に示す．正規化を行い最も認識率が高かった結果は正規化を行わなかったときの結果を比較するとどの言語においても認識率が向上しており，平均で 2.5 ポイント向上した．よって，どの言語においても正規化することで発話内容による特徴量の変化を吸収できることが分かった．また言語ごとに認識しやすい言語とそうでない言語があり，例えば正規化を行った場合でも英語は 41.3%であるのに対してイタリア語は 86.7%と認識率に大きな差が生じた．

3.6 既知言語に対する実験と結果

識別器を多言語化できると識別器を言語ごとに作成せず言語を混ぜて 1 つにすることが可能である．そこでまず既知言語に対する性能評価，つまり評価対象言語を学習言語

に含むときに感情識別が可能かどうか検証する．学習に使用する言語数を変化させ，認識精度がどう変化するか実験を行った．実験結果を表 3 に示す．ここでの「1 言語学習」は 2 と同じである．また「2 言語学習」の場合，自身の言語と他の 1 言語を合わせて学習して言語の組み合わせを変えて 3 回実験し，その平均を明記している．これを見るとどの言語においても正規化を行った場合の方が認識率が高いことが分かる．しかし正規化を行った場合でも 4 言語で学習した場合は 1 言語で学習した場合と比較すると平均で 4.3 ポイント認識率が低下した．これは自身の言語の情報が減ってしまうためだと考えられる．しかし評価対象言語のみで正規化を行わない場合の認識率は平均して 60.6%，一方 4 言語で正規化ありで学習を行った場合は平均して 58.9%と両者にあまり差がないことから言語を混ぜて識別器を作成しても正規化を行えばほとんど性能劣化がないということが分かった．

3.7 未知言語に対する実験と結果

識別器を多言語化できると未知の言語であっても感情識別を行うことができる．そこで評価対象言語を学習言語に含まないときの認識精度を検証して未知言語での識別が可能かどうか実験を行った．実験結果を表 4 に記す．1 言語学習のとき，既知言語で評価を行った場合と比較すると，例えばイタリア語では正規化なしで 85.2%から 24.0%，正規化ありで 86.7%から 29.0%と認識率はかなり低下していることが分かる．このように自身の言語を含まずに 1 言語で学習した場合は自身の言語のみを用いて学習を行う場合より認識率がかなり下がってしまうということがどの言語においても言える．また学習する言語の数が増えても認識率が変化しないということが分かる．よって自身の言語を学習に使用しない場合は他の言語の情報が増えても影響はほとんどないということが言える．

3.8 正規化の効果

既知言語と未知言語の性能比較を行った．正規化を行った場合の認識率を言語ごとに算出して 4 言語での平均を計算すると既知言語の場合は 58.9%，未知言語の場合は 3 言語で 30.2%となり，未知言語での認識率はかなり低くなってしまうことが分かる．このことから正規化を行った場合でも言語間の違いは完全には吸収できず，未知言語において感情識別を行うのは厳しいということが言える．また正規化を行わないとき既知言語の場合は 57.5%，未知言語の場合は 24.0%となった．正規化の効果を調べるため，正規化を行う場合と行わない場合において対象言語が既知から未知になることによる識別率の減少率を計算したところ，正規化ありで 48.7%，正規化なしで 58.3%となり正規化を行う場合の方が減少率が小さいという結果になった．このことから，正規化を行うことで言語間の違いを多少ではあるが吸収できると言える．

さらに正規化を行うことによる認識率の改善率を計算した．学習言語数ごとに正規化を行った場合，どれほど認識率が向上するか改善率を計算してそれらの平均を言語ごとに算出した結果を表 5 に示す．この結果から自身の言語を学習に使用しない場合は使用する場合と比べて改善幅が大きいことが分かる．これによっても正規化を行うことで言語の違いを多少は吸収できることが分かる．

4. まとめ

自分が知らない言語でも音声から感情は大体読み取ることができる．そこで感情識別器を多言語化できるのではないかと考えた．感情識別器の多言語化が可能になると識別器を 1 つにできるということと未知言語による感情識別が可能になるといったメリットがある．しかし言語によってアクセントやイントネーションなどが違うため，同じ感情でも声の高さや大きさなどの特徴量が異なるため，そのまま言語を混ぜて識別器を作成しても感情識別が上手く行えなくなるからである．そこで感情を含む場合にが感情の含まない場合と比べて特徴量がどれだけ変化しているかに注目して，正規化を行うことによって言語間の違いを取り除くことができるのではないかと仮定した．その検証を行うために，日本語，ドイツ語，英語，イタリア語の 4 言語を用いて実験を行った．

言語ごとに認識精度を算出した結果，正規化を用いる場合の方が認識率が高くなり，正規化によって発話内容による違いを吸収することができるということが分かった．また自身の言語を学習に含む場合は正規化を行っても自身の言語の情報が減ってしまうため，学習言語数を増やすごとに認識率が低下してしまうという結果になった．しかし，正規化を用いずに自身の言語のみで学習を行った場合の認識率は平均して 60.6%，正規化を用いて 4 言語で学習を行った場合の認識率は平均して 58.9%となり，双方にあまり差がないことから既知言語で感情識別を行う際は単一の識別器を用いることができることが分かった．次に自身を学習に含まない場合で実験を行うことで未知言語の識別が可能かどうか調査した．結果，正規化を用いて自身のみで学習を行った場合 (4 言語学習) の認識率は平均して 58.9%，自身を学習に使用しない場合 (3 言語学習) では認識率は平均して 30.2%となり，未知言語で識別を行った場合はかなり認識率が低下した．このことから言語間の違いは正規化を用いても完全には取り除くことができないことが分かる．しかし正規化を行わない場合と行う場合での変化量に注目すると，既知言語での識別では平均して 4.6%，未知言語での識別では平均して 27.7%となり，未知言語で識別を行った方が改善幅が大きいため，多少ではあるが正規化を用いることで言語間の違いを取り除くことができるということが分かる．

表 3 評価言語を学習に含む場合の識別結果

認識対象言語	日本語		ドイツ語		英語		イタリア語		平均	
正規化の有無	なし	あり	なし	あり	なし	あり	なし	あり	なし	あり
1 言語学習	48.6 %	49.5 %	73.9 %	74.9 %	34.7 %	41.3 %	85.2 %	86.7 %	60.6 %	63.1 %
2 言語学習	49.4 %	49.7 %	62.9 %	70.9 %	34.4 %	39.0 %	80.8 %	82.1 %	56.9 %	60.4 %
3 言語学習	48.3 %	49.2 %	67.6 %	69.3 %	34.1 %	38.4 %	78.8 %	80.5 %	57.2 %	59.4 %
4 言語学習	48.2 %	49.3 %	66.7 %	69.1 %	36.0 %	38.0 %	79.0 %	79.0 %	57.5 %	58.9 %

表 4 評価言語を学習に含まない場合の識別結果

認識対象言語	日本語		ドイツ語		英語		イタリア語		平均	
正規化の有無	なし	あり	なし	あり	なし	あり	なし	あり	なし	あり
1 言語学習	20.8 %	25.3 %	25.0 %	36.0 %	20.4 %	26.1 %	24.0 %	29.0 %	22.6 %	29.1 %
2 言語学習	20.9 %	23.9 %	27.8 %	41.3 %	19.4 %	25.6 %	22.9 %	28.7 %	22.8 %	29.9 %
3 言語学習	22.4 %	24.8 %	28.8 %	42.1 %	20.3 %	25.7 %	24.5 %	28.1 %	24.0 %	30.2 %

表 5 正規化を用いることによる改善率

	既知言語	未知言語
日本語	1.7 %	15.6 %
ドイツ語	2.5 %	46.2%
英語	12.6 %	28.8 %
イタリア語	1.4 %	20.3 %

参考文献

[1] Apple(日本)：Siri, https://www.apple.com/jp/ios/siri.
[2] Google: Goole Home, https://store.google.com/product/google$_h ome$.
[3] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, Stefan Scherer: Representation Learning for Speech Emotion Recognition, INTERSPEECH, ISCA, San Francisco, USA, pp. 3603-3607 (2016).
[4] Aharon Satt, Shai Rozenberg, Ron Hoory: Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms, INTERSPEECH, ISCA, Stockholm, Sweden, pp. 1089-1093 (2017).
[5] 中川祥平：合成音声を用いた特徴量の正規化による感情識別法，電子情報通信学会論文誌 D Vol.J97-D No.3 pp.533-539 (2014).
[6] 有本泰子, 河津宏美：感情評定値付きオンラインゲーム音声チャットコーパス (OGVC) 使用説明書, 人工知能学会論文誌, Vol.29, No.1SP1-B, pp.11-20 (2014).
[7] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter Sendlmeier und Benjamin Weiss: A Database of German Emotional Speech, 入手先 ⟨http://emodb.bilderbar.info/start.html⟩ (2005).
[8] Philip Jackson and Sana-UI Haq: Surrey Aidio-Visual Expressed Emotion (SAVEE) Database, 入手先 ⟨http://kahlan.eps.surrey.ac.uk/savee/⟩ (2010).
[9] Giovanni Costantini, Iacopo Iadarola, Andrea Paoloni, Massimiliano Todisco: EMOVO Corpus: an Italian Emotional Speech Databas, 入手先 ⟨http://voice.fub.it/EMOVO⟩ (2014).
[10] 名古屋工業大学：Open JTalk, 入手先 ⟨http://open-jtalk.sourceforge.net⟩.
[11] Smart Link Corporation: Text To Speech, 入手先 ⟨http://text-to-speech.imtranslator.net/speech.asp⟩.
[12] Bjorn Schuller, Stefan Steidl, and Anton Batliner: The INTERSPEECH 2009 Emotion Challenge, Interspeech 2009 Speech and Intelligence, pp.312-315 (2009).
[13] Chih-Chung Chang, Chih-Jen Lin: LIBSVM: A Library for Support Vector Machines, 入手先 ⟨http://www.csie.ntu.edu.tw⟩ (2001).