Multilingual Speech Emotion Recognition on Japanese, English, and German

Panikos Heracleous[1], Keiji Yasuda[1,2], Akio Yoneyama[1]

[1]KDDI Research, Inc.
Education and Medical ICT Laboratory
2-1-15 Ohara, Fujimino-shi, Saitama, 356-8502, Japan
{*pa-heracleous,yoneyama*}*@kddi-research.jp*
[2]Nara Institute of Science and Technology, Japan
ke-yasuda@dsc.naist.jp

Abstract. The current study focuses on human emotion recognition based on speech, and particularly on multilingual speech emotion recognition using Japanese, English, and German emotional corpora. The proposed method exploits conditional random fields (CRF) classifiers in a two-level classification scheme. Specifically, in the first level, the language spoken is identified, and in the second level, speech emotion recognition is carried out using emotion models specific to the identified language. In both the first and second levels, CRF classifiers fed with acoustic features are applied. The CRF classifier is a popular probabilistic method for structured prediction, and is widely applied in natural language processing, computer vision, and bioinformatics. In the current study, the use of CRF in speech emotion recognition when limited training data are available is experimentally investigated. The results obtained show the effectiveness of using CRF when only a small amount of training data are available and methods based on a deep neural networks (DNN) are less effective. Furthermore, the proposed method is also compared with two popular classifiers, namely, support vector machines (SVM), and probabilistic linear discriminant analysis (PLDA) and higher accuracy was obtained using the proposed method. For the classification of four emotions (i.e., neutral, happy, angry, sad) the proposed method based on CRF achieved classification rates of 93.8% for English, 95.0% for German, and 88.8% for Japanese. These results are very promising, and superior to the results obtained in other similar studies on multilingual or even monolingual speech emotion recognition.

Keywords: Speech emotion recognition, multilingual, conditional random fields, two-level classification, i-vector paradigm, deep learning

1 Introduction

Automatic recognition of human emotions [1] is a relatively new field, and is attracting considerable attention in research and development areas because of

its high importance in real applications. Emotion recognition can be used in human-robot communication, when robots communicate with humans according to the detected human emotions, and also has an important role at call centers to detect the caller's emotional state in cases of emergency (e.g., hospitals, police stations), or to identify the level of a customer's satisfaction (i.e., providing feedback). In the current study, multilingual emotion recognition based on speech is experimentally investigated. Specifically, using English, German, and Japanese emotional speech data, multilingual emotion recognition experiments are conducted based on several classification approaches and the i-vector paradigm framework.

Previous studies reported automatic speech emotion recognition using Gaussian mixture models (GMMs) [2], support vector machines [3], neural networks (NN) [4], and deep neural networks (DNN) [5]. Most studies in speech emotion recognition have focused solely on a single language, and cross-corpus speech emotion recognition has been addressed in only a few studies. In [6], experiments on emotion recognition are described using comparable speech corpora collected from American English and German interactive voice response systems, and the optimal set of acoustic and prosodic features for mono-, cross-, and multilingual anger recognition are computed. Cross-language speech emotion recognition based on HMMs and GMMs is reported in [7]. Four speech databases for cross-corpus classification, with realistic, non-prompted emotions and a large acoustic feature vector are reported in [8].

In the current study, however, multilingual speech emotion recognition using Japanese, English, and German corpora based on a two-level classification scheme is demonstrated. Specifically, spoken language identification and emotion recognition are integrated in a complete system capable of recognizing four emotions from English, German, and Japanese databases. In the first level, spoken language identification using emotional speech is performed, and in the second level the emotions are classified using acoustic models of the language identified in the first level. For classification in both the first and second levels, CRF classifiers are applied and compared to SVM and PLDA classifiers. A similar study –but with different objectives– is presented in [9]. In a more recent study [10], a three-layer perception model is used for multilingual speech emotion recognition using Japanese, Chinese, and German emotional corpora. In that specific study, the volume of training and test data used in classification is closely comparable with the data used in the current study, and, therefore, comparisons are, to some extent, possible. Although very limited training data were available, DNN and convolutional neural networks (CNN) were also considered for comparison purposes.

Automatic language identification (LID) is a process whereby a spoken language is identified automatically. Applications of language identification include, but are not limited to, speech-to-speech translation systems, re-routing incoming calls to native speaker operators at call centers, and speaker diarization. Because of the importance of spoken language identification in real applications, many studies have addressed this issue. The approaches reported are categorized

into the acoustic-phonetic approach, the phonotactic approach, the prosodic approach, and the lexical approach [11]. In phonotactic systems [11, 12], sequences of recognized phonemes obtained from phone recognizers are modeled. In [13], a typical phonotactic language identification system is used, where a language dependent phone recognizer is followed by parallel language models (PRLM). In [14], a universal acoustic characterization approach to spoken language recognition is proposed. Another method based on vector-space modeling is reported in [11, 15], and presented in [16].

In acoustic modeling-based systems, different features are used to model each language. Earlier language identification studies reported methods based on neural networks [17, 18]. Later, the first attempt at using deep learning has also been reported [19]. Deep neural networks for language identification were used in [20]. The method was compared with i-vector-based classification, linear logistic regression, linear discriminant analysis-based (LDA), and Gaussian modeling-based classifiers. In the case of a large amount of training data, the method demonstrated its superior performance. When limited training data were used, the i-vector yields the best identification rate. In [21] a comparative study on spoken language identification using deep neural networks was presented. Other methods based on DNN and recurrent neural networks (RNN) were presented in [22, 23]. In [24], experiments on language identification using i-vectors and CRF were reported. The i-vector paradigm for language identification with SVM [25] was also applied in [26]. SVM with local Fisher discriminant analysis is used in [27]. Although significant improvements in LID have been achieved using phonotactic approaches, most state-of-the-art systems still rely on acoustic modeling.

2 Methods

2.1 Emotional Speech Data

Four professional female actors simulated Japanese emotional speech. These comprised neutral, happy, angry, sad, and mixed emotional states. Fifty-one utterances for each emotion were produced by each speaker. The sentences were selected from a Japanese book for children. The data were recorded at 48 kHz and down-sampled to 16 kHz, and they also contained short and longer utterances varying from 1.5 sec to 9 sec. Twenty-eight utterances were used for training and 20 for testing. The remaining utterances were excluded due to poor speech quality.

For the English emotional speech data, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) set [28] was used. RAVDESS uses a set of 24 actors (12 male, 12 female) speaking and singing with various emotions, in a North American English accent, and contains 7,356 high-quality video recordings of emotionally-neutral statements, spoken and sung with a range of emotions. The speech set consists of the 8 emotional expressions: neutral, calm, happy, sad, angry, fearful, surprised, and disgusted. The song set consists of the 6 emotional expressions: neutral, calm, happy, sad, angry, and fearful. All emotions except

neutral are expressed at two levels of emotional intensity: normal and strong. There are 2,452 unique vocalizations, all of which are available in three formats: full audio-video (720p, H.264), video only, and audio only (wav). The database has been validated in a perceptual experiment involving 297 participants. The data are encoded as 16-bit, 48-kHz wav files, and down-sampled to 16 kHz. In the current study, 96 utterances for neutral, happy, angry, and sad emotional states were used. For training, 64 utterances were used for each emotion, and 32 for testing.

The German database used was the Berlin database [29], which includes seven emotional states: anger, boredom, disgust, anxiety, happiness, sadness, and neutral speech. The utterances were produced by ten professional German actors (5 female, 5 male) speaking ten sentences with an emotionally neutral content but expressed with the seven different emotions. The actors produced 69 frightened, 46 disgusted, 71 happy, 81 bored, 79 neutral, 62 sad, and 127 angry emotional sentences. For training, 42 utterances were used in the study, and for testing, 20 utterances, in the neutral, happy, angry, and sad modes.

2.2 Classification Approaches

Conditional Random Fields (CRF) CRF is a modern approach similar to HMMs, however with a different nature. CRF are undirected graphical models, a special case of conditionally trained finite state machines. They are discriminative models, which maximize the conditional probability of observation and state sequences. CRF assume frame dependence, and as a result context is also considered. The main advantage of CRF is their flexibility to include a wide variety of non-independent features. CRF have been successfully used for meeting segmentation [30], for phone classification [31], and for events recognition and classification [32]. A language identification method based on deep-structured CRF has been reported in [33]. The current study is based on the popular and very simple linear-chain CRF, along with low dimensional feature representation using i-vectors. Similarly, to [34] for object recognition using CRF, each input sentence in represented by a single vector (i.e., an i-vector), and this scenario is different from the conventional classification approaches in machine learning, where the input space is represented as a set of feature vectors.

In CRF, the probabilities of a class label s given the observation sequence $\mathbf{o} = (o_1, o_2, ..., o_T)$ are given by the following equation:

$$p(k|\mathbf{o}, \lambda) = \frac{1}{z(\mathbf{o}, \lambda)} \sum_{\mathbf{s} \in k} e^{\lambda \cdot f(k, \mathbf{s}, \mathbf{o})} \tag{1}$$

where λ is the parameter vector, $\mathbf{f}$ is the sufficient statistics vector, and $\mathbf{s} = (s_1, s_2, ..., s_T)$ is a hidden state sequence. The function $z(\mathbf{o}, \lambda)$ ensures that the model forms a properly normalized probability and is defined as:

$$z(\mathbf{o}, \lambda) = \sum_{k} \sum_{\mathbf{s} \in k} e^{\lambda \cdot f(k, \mathbf{s}, \mathbf{o})} \tag{2}$$

Figure 1 demonstrates the structure of HMM and CRF models.

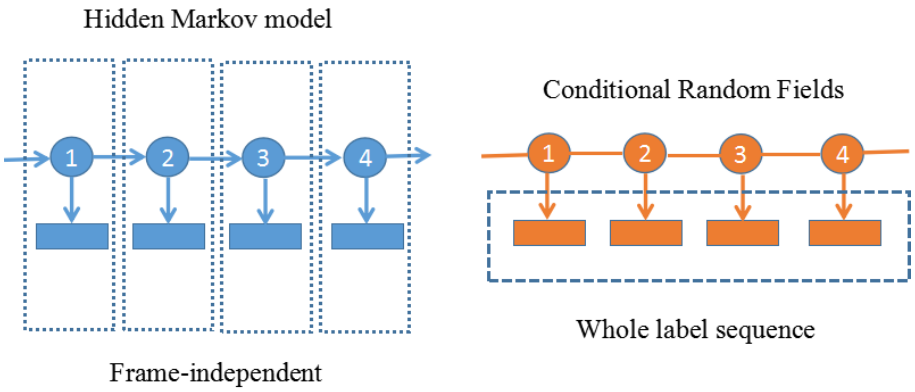


Fig. 1: Structures of hidden Markov models (HMM) and conditional random fields (CRF).

Support Vector Machines (SVM) A support vector machine (SVM) is a two-class classifier constructed from sums of a kernel function *K(.,.)*

$$f(x) = \sum_{i=1}^{L} \alpha_i t_i K(\mathbf{x}, \mathbf{x_i}) + d \tag{3}$$

where the t_i are the ideal outputs, $\sum_{i=1}^{L} \alpha_i t_i = 0$, and $\alpha_i > 0$.

An SVM is a discriminative classifier, which is widely used in regression and classification. Given a set of labeled training samples, the algorithm finds the optimal hyperplane, which categorizes new samples. SVM is among the most popular machine learning methods. The advantages of SVM include the support of high-dimensionality, memory efficiency, and versatility. However, when the number of features exceeds the number of samples the SVM performs poorly. Another disadvantage is that SVM is not probabilistic because it works by categorizing objects based on the optimal hyperplane.

Originally, SVMs were used for binary classification. Currently, the multi-class SVM, a variant of the conventional SVM, is widely used in solving multi-class classification problems. The most common way to build a multi-class SVM is to use K one-versus-rest binary classifiers (commonly referred to as "one-versus-all" or OVA classification). Another strategy is to build one-versus-one classifiers, and to choose the class that is selected by the most classifiers. In this case, *K(K-1)/2* classifiers are required and the training time decreases because less training data are used for each classifier.

Probabilistic Linear Discriminant Analysis (PLDA) PLDA is a popular technique for dimension reduction using the Fisher criterion. Using PLDA, new axes are found, which maximize the discrimination between the different classes.

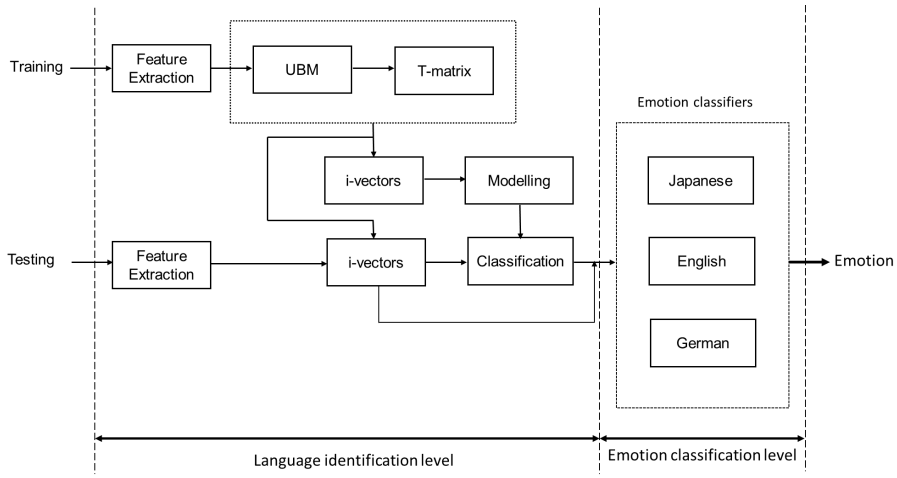


Fig. 2: Classification scheme based on the i-vector paradigm.

PLDA was originally applied to face recognition, and and can be used to specify a generative model of the i- vector representation. A study using UBM-based LDA for speaker recognition was also presented in [35]. Adapting this to language identification and emotion classification, for the i-th language or emotion, the i-vector $\mathbf{w}_{i,j}$ representing the j-th recording can be formulated as:

$$\mathbf{w}_{i,j} = \mathbf{m} + \mathbf{S}\mathbf{x}_i + \mathbf{e}_{i,j} \tag{4}$$

where $\mathbf{S}$ represents the between-language or between-emotion variability, and the latent variable $\mathbf{x}$ is assumed to have a standard normal distribution, and to represent a particular language or emotion and channel. The residual term $\mathbf{e}_{i,j}$ represents the within-language or within-emotion variability, and it is assumed to have a normal distribution. Figure 2 shows the two-level classification scheme used in the current study.

2.3 Shifted Delta Cepstral (SDC) Coefficients

Previous studies showed that language identification performance is improved by using SDC feature vectors, which are obtained by concatenating delta cepstra across multiple frames. The SDC features are described by the N number of cepstral coefficients, d time advance and delay, k number of blocks concatenated for the feature vector, and P time shift between consecutive blocks. For each SDC final feature vector, kN parameters are used. In contrast, in the case of conventional cepstra and delta cepstra feature vectors, $2N$ parameters are used. The SDC is calculated as follows:

$$\Delta c(t + iP) = c(t + iP + d) - c(t + iP - d) \tag{5}$$

The final vector at time t is given by the concatenation of all $\Delta c(t + iP)$ for all $0 \leq i < k$, where $c(t)$ is the original feature value at time t. In the current study, SDC coefficients were used not only in spoken language identification, but also in emotion classification.

2.4 Feature extraction

In automatic speech recognition, speaker recognition, and language identification, mel-frequency cepstral coefficients (MFCC) are among the most popular and most widely used acoustic features. Therefore, in modeling the languages being identified and the emotions being recognized, this study similarly used 12 MFCC, concatenated with SDC coefficients to form feature vectors of length 112. The MFCC features were extracted every 10 ms using a window-length of 20 ms. The extracted acoustic features were used to construct the i-vectors used in emotion and spoken language identification modeling and classification.

A widely used approach for speaker recognition is based on Gaussian mixture models (GMM) with universal background models (UBM). The individual speaker models are created using maximum a posteriori (MAP) adaptation of the UBM. In many studies, GMM supervectors are used as features. The GMM supervectors are extracted by concatenating the means of the adapted model.

The problem of using GMM supervectors is their high dimensionality. To address this issue, the i-vector paradigm was introduced which overcomes the limitations of high dimensionality. In the case of i-vectors, the variability contained in the GMM supervectors is modeled with a small number of factors, and the whole utterance is represented by a low dimensional i-vector of 100-400 dimension.

Considering language identification, an input utterance can be modeled as:

$$\mathbf{M} = \mathbf{m} + \mathbf{Tw} \tag{6}$$

where $\mathbf{M}$ is the language-dependent supervector, $\mathbf{m}$ is the language-independent supervector, $\mathbf{T}$ is the total variability matrix, and $\mathbf{w}$ is the i-vector. Both the total variability matrix and language-independent supervector are estimated from the complete set of the training data. The same procedure is used to extract i-vectors used in speech emotion recognition.

2.5 Evaluation measures

In the current study, the equal error rates (EER) (i.e., equal false alarms and miss probability) and the classification rates are used as evaluation measures. The classification rate is defined as:

$$acc = \frac{1}{n} \sum_{k=1}^{n} \frac{No.\ of\ corrects\ for\ class\ k}{No.\ of\ trials\ for\ class\ k} \cdot 100 \tag{7}$$

where n is the number of the emotions. In addition, the detection error trade-off (DET) curves, which show the function of miss probability and false alarms, are also given.

3 Results

This section presents the results for multilingual emotion classification based on a two-level classification scheme using Japanese, English, and German corpora.

3.1 Spoken language identification using emotional speech data

The i-vectors used in modeling and classification are constructed using MFCC features and SDC coefficients. For training, 160 utterances from each language are used, and 80 utterances for testing. The dimension of the i-vectors is set to 100, and 256 Gaussian components are used in the UBM-GMM. Due to the fact that only three target languages are used, the identification was perfect almost in all cases (except in the case of using PLDA was 98.8%). On the other hand, it should be noted that language identification is conducted using emotional speech data, and this result indicates that spoken language classification using emotional speech data does not present any particular difficulties compared to normal speech.

3.2 Emotion recognition based on a two-level classification scheme

Table 1 shows the average emotion classification rates when using MFCC features only. As shown, high classification rates are being obtained. The results show that the two classifiers based on DNN and CNN show lower rates (except for Japanese). A possible reason may be the small volume of training data in the case of English and German.

Table 2 shows the average classification rates when using MFCC features along with SDC coefficients. As shown, the CRF classifier shows superior performance in most of cases, followed by SVM. The results show that using SDC coefficients along with MFCC features improves classification rates. This result indicates that SDC coefficients are effective not only in spoken language identification, but also in speech emotion recognition. Note, however, that in this case of DNN and CNN, small or no improvements are being obtained. The results indicate that due to the limited training data, DNN and CNN are less effective for this task.

Table 3 shows the classification rates for the four emotions when using the CRF classifier and MFCC features along with SDC coefficients. In the case of Japanese the average accuracy was 88.8%, in the case of English the average was 93.8%, and in the case of German, a 95.0% accuracy was obtained. Concerning the German corpus, the results obtained are significantly higher compared to the results reported in [36] when the same corpus was used.

Table 4 shows the individual classification rates when SVM was used. In the case of Japanese, a 82.8% average accuracy was achieved, in the case of English the average accuracy was 91.4%, and when using the German corpus the average accuracy was 95.0%.

Table 5 shows the recognition rates when using the PLDA classifier. The average accuracy for Japanese was 85.2%, the accuracy for English was 90.2%,

Table 1: Average emotion classification rates when using MFCC features for the i-vector construction.

Classifier	*Language*		
	Japanese	English	German
PLDA	85.2	77.3	**91.7**
CRF	79.4	**87.5**	90.0
SVM	82.8	80.5	91.3
DNN	**90.6**	68.3	85.2
CNN	90.2	71.0	88.7

Table 2: Average emotion classification rates when using MFCC features and SDC coefficients for the i-vector construction.

Classifier	*Corpus*		
	Japanese	English	German
PLDA	87.6	90.9	91.7
CRF	88.8	**93.8**	**95.0**
SVM	**90.9**	93.0	**95.0**
DNN	83.7	76.2	82.7
CNN	88.8	77.1	84.5

and for the German corpus an accuracy of 89.3% was achieved. The results show that when using CRF, superior performance was obtained, followed by SVM. The lowest rates were obtained when the PLDA classifier was used. The results also show that the emotion *sad* is recognized with the highest rates in most cases.

3.3 Emotion recognition using multilingual emotion models

In this baseline approach, a single-level classification scheme is used. Using emotional speech data from Japanese, English, and German languages, common emotion models are trained. For training, 112 Japanese, 64 English, and 40 German i-vectors are used for each emotion. For testing, 80 Japanese, 32 English, and 20 German i-vectors are used for each emotion. Since also using SDC coefficients improves the performance of the two-level approach, in this method, i-vectors are constructed using MFCC features in conjunction with SDC coefficients. Table 6 shows the classification rates. As shown, using a universal multilingual model, the average emotion classification accuracies for the three languages are 75.2%, 77.7%, and 75.0% when using PLDA, CRF, and SVM classifiers, respectively. This is a promising result and superior to the results obtained in other similar studies. While the rates achieved are lower than with the two-level approach, in this approach a single level is used with reduced system complexity (i.e., language identification is not applied). Furthermore, the classification rates may be improved with a larger amount of training data. These result show that

Table 3: Emotion classification rates when using CRF classifier and MFCC features, along with SDC coefficients, for the i-vector construction.

Corpus	*Emotions*				
	Neutral	Happy	Anger	Sad	Average
Japanese	85.0	83.8	88.8	**97.5**	88.8
English	87.5	**100.0**	96.9	90.6	93.8
German	**100.0**	95.0	85.0	100.0	95.0

Table 4: Emotion classification rates when using SVM classifier and MFCC features, along with SDC coefficients, for the i-vector construction.

Corpus	*Emotions*				
	Neutral	Happy	Anger	Sad	Average
Japanese	**92.5**	70.0	81.3	87.5	82.8
English	84.4	**100.0**	93.8	87.5	91.4
German	95.0	95.0	90.0	**100.0**	95.0

Table 5: Emotion classification rates when using PLDA classifier and MFCC features, along with SDC coefficients, for the i-vector construction.

Corpus	*Emotions*				
	Neutral	Happy	Anger	Sad	Average
Japanese	72.8	86.4	90.1	**91.4**	85.2
English	93.9	90.9	**97.0**	78.8	90.2
German	**95.2**	81.0	85.7	**95.2**	89.3

Table 6: Average emotion classification rates when using a universal emotion model with MFCC features and SDC coefficients for the i-vector construction.

Classifier	*Corpus*			
	Japanese	English	German	Average
PLDA	75.3	68.2	**82.1**	75.2
CRF	**80.9**	**73.4**	78.8	**77.7**
SVM	76.9	71.9	76.3	75.0

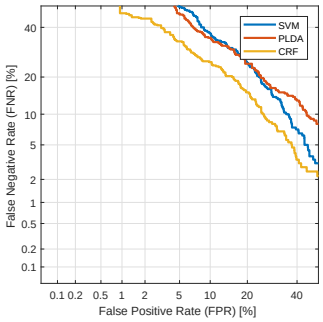
i-vectors can efficiently be applied in multilingual emotion recognition when universal, multilingual emotion models are also used. The results also show that in most cases, the performance of the CRF classifier is superior.

Table 7 shows the EER when a universal, multilingual emotion model is used. As shown, the EER for German is the lowest among the three, followed by the EER for Japanese. The average EERs for the three languages are 20.9%, 19.3%, and 23.2% when using PLDA, CRF, and SVM classifiers, respectively. Also in

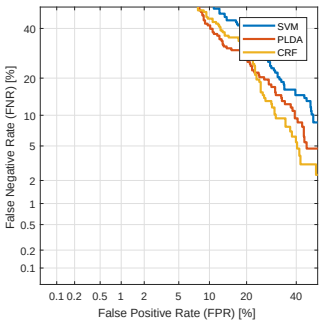
Table 7: Equal error rates (EER) when using a universal emotion model with MFCC features and SDC coefficients for the i-vector construction.

Classifier	*Corpus*			
	Japanese	English	German	Average
PLDA	22.6	**22.6**	**17.5**	20.9
CRF	**17.6**	22.9	**17.5**	**19.3**
SVM	22.5	26.8	20.4	23.2

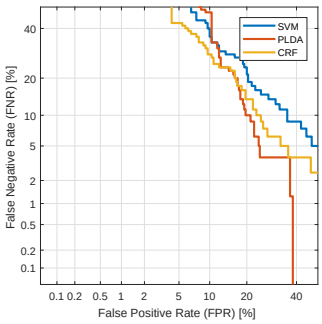
this case, the lowest EERs were obtained using the CRF classifier. Figure 3 shows the DET curves for multilingual emotion recognition using a universal emotion model.



(a) Japanese



(b) (English



(c) German

Fig. 3: DET curves for the three languages used in emotion classification when using common multilingual models.

4 Discussion

Although using real-world emotional speech data would represent a more realistic situation, acted emotional speech data are widely used in speech emotion classification. Furthermore, the current study mainly investigated classification schemes and features extraction methods, so using acted speech is a reasonable and acceptable approach. Because of limited emotional data, deep learning approaches in multilingual emotion recognition were not investigated. In contrast, a method is proposed that integrates spoken language identification and emotion classification. In addition to SVM and PLDA classifiers, the CRF classifier is also used in combination with the i-vector paradigm. The results obtained show the advantage of using the CRF classifier, especially when limited data are available. For comparison purposes, deep neural networks were also considered. Because of the limited training data, however, the classification rates when using DNN and CNN were significantly lower. In order to address the problems associated with using acted speech, an initiative to obtain a large quantity of spontaneous emotional speech is currently being undertaken. With such data, it will also be possible to analyze the behavior of additional classifiers, such as deep neural networks, and to investigate the problem of multilingual speech emotion recognition in realistic situations (e.g., noisy or reverberant environments).

5 Conclusions

The current study experimentally investigated multilingual speech emotion classification. A two-level classification approach was used, integrating spoken language identification and emotion recognition. The proposed method was based on CRF classifier and the i-vector paradigm. When classifying four emotions, the proposed method achieved a 93.8% classification rate for English, a 95.0% rate for German, and 88.8% rate for Japanese. These results were very promising, and demonstrated the effectiveness of the proposed methods in multilingual speech emotion recognition. An initiative to obtain realistic, spontaneous emotional speech data for a large number of languages is currently being undertaken. As future work, the effect of noise and reverberation will also be investigated.

References

1. C. Busso, M. Bulut, and S. Narayanan, "Toward Effective Automatic Recognition Systems of Emotion in Speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford University Press, November 2013, pp. 110–127.
2. H. Tang, S. Chu, and M. H. Johnson, "Emotion Recognition From Speech Via Boosted Gaussian Mixture Models," *in Proc. of ICME*, pp. 294–297, 2009.
3. Y. Pan, P. Shen, and L. Shen, "Speech Emotion Recognition Using Support Vector Machine," *International Journal on Smart Home*, vol. 6 (2), pp. 101–108, 2012.

4. J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion Recognition in Speech Using Neural Networks," *Neural Computing & Applications*, vol. 9, Issue 4, pp. 290–296, 2000.
5. K. Han, D.Yu, and I.Tashev, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine," *in Proc. of Interspeech*, pp. 223–227, 2014.
6. T. Polzehl, A. Schmitt, , and F. Metze, " Approaching multi-lingual emotion recognition from speech-on language dependency of acoustic prosodic features for anger detection," *in Proc. of Speech Prosody*, 2010.
7. M. Bhaykar, J. Yadav, and K. S. Rao, "Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM," *in Communications (NCC), 2013 National Conference on. IEEE*, pp. 1–5, 2013.
8. F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Crosscorpus classification of realistic emotions - some pilot experiments," *in Proc. of the Third International Workshop on EMOTION (satellite of LREC)*, 2010.
9. H. Sagha, P. Matejka, M. Gavryukova, F. Povolny, E. Marchi, and B. Schuller, " Enhancing Multilingual Recognition of Emotion in Speech by Language Identification," *in Proc. of Interspeech*, 2016.
10. X. Li and M. Akagi, "A Three-Layer Emotion Perception Model for Valence and Arousal-Based Detection from Multilingual Speech," *in Prof. of Interspeech*, pp. 3643–3647, 2018.
11. H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: From fundamentals to practice," *in Proc. of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.
12. M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *lEEE Transactions on Speech and Audio Processing*, vol. 4(1), pp. 31–44, 1996.
13. D. Caseiro and I. Trancoso, "Spoken language identification using the speechdat corpus," *In Proc. of ICSLP'98*, 1998.
14. S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer speech and language*, vol. 27, pp. 209–227, 2013.
15. C.-H. Lee, "Principles of spoken language recognition," *in Springer Handbook on Speech Processing and Speech Communication, J. Benesty, Y. Hunag. M. M. Sondhi, Editors, SpringerVerlag*, 2008.
16. D. A. Reynolds, W. M. Campbell, W. Shen, and E. Singer, "Automatic language recognition via spectral and token based approaches," *in Springer Handbook on Speech Processing and Speech Communication, J. Benesty, Y. Hunag. M. M. Sondhi, Editors, SpringerVerlag*, 2008.
17. R. Cole, J. Inouye, Y. Muthusamy, and M. Gopalakrishnan, "Language identification with neural networks: a feasibility study," *in Proc. of IEEE Pacific Rim Conference*, pp. 525–529, 1989.
18. M. Leena, K. S. Rao, and B. Yegnanarayana, "Neural network classifiers for language identification using phonotactic and prosodic features," *in Proc. of Intelligent Sensing and Information Processing*, pp. 404–408, 2005.
19. G. Montavon, "Deep learning for spoken language identification," *in NIPS workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.
20. I. L.-Moreno, J. G.-Dominguez, O. Plchot, D. Martinez, J. G.-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," *in Proc. of ICASSP*, pp. 5337–5341, 2014.

21. P. Heracleous, K. Takai, K. Yasuda, Y. Mohammad, and A. Yoneyama, "Comparative Study on Spoken Language Identification Based on Deep Learning ," *in Proc. of EUSIPCO*, 2018.
22. B. Jiang, Y. Song, S. Wei, J.-H. Liu, I. V.McLoughlin, and L.-R. Dai, "Deep bottleneck features for spoken language identification," *PLos ONE*, vol. 9(7), pp. 1–11, 2010.
23. R. Zazo, A. L.-Diez, J. G.-Dominguez, D. T. Toledano, and J. G.-Rodriguez, "Language identification in short utterances using long short-term memory (lstm) recurrent neural networks," *PLos ONE*, vol. 11(1): e0146917., 2016.
24. P. Heracleous, Y. Mohammad, K. Takai, K. Yasuda, and A. Yoneyama, "Spoken Language Identification Based on I-vectors and Conditional Random Fields," *in Proc. of IWCMC*, pp. 1443–1447, 2018.
25. N. Cristianini and J. S.-Taylor, "Support vector machines," *Cambridge University Press, Cambridge*, 2000.
26. N. Dehak, P. A.T.-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via ivectors and dimensionality reduction," *in Proc. of Interspeech*, pp. 857–860, 2011.
27. P. Shen, X. Lu, L. Liu, and H. Kawai, "Local fisher discriminant analysis for spoken language identification," *in Proc. of ICASSP*, pp. 5825–5829, 2016.
28. S. R. Livingstone, K. Peck, F.A., and Russo, " RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song," *in 22nd Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science (CSBBCS)(Kingston, ON.).*, 2012.
29. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," *in Proc. of Interspeech*, 2005.
30. S. Reiter, B. Schuller, and G. Rigoll, "Hidden Conditional Random Fields for Meeting Segmentation," *in Proc. of ICME*, pp. 639–642, 2007.
31. A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden Conditional Random Fields for Phone Classification," *in Proc. of Interspeech*, pp. 1117–1120, 2005.
32. H. Llorens, E. Saquete, and B. N.-Colorado, "TimeML Events Recognition and Classification: Learning CRF Models with Semantic Roles," *in Proc. of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 725–733, 2010.
33. D. Yu, S. Wang, Z. Karam, and L. Deng, "Language Recognition Using Deep-structured Conditional Random Fields," *in Proc. of ICASSP*, pp. 5030–5033, 2010.
34. A. Quattoni, M. Collins, and T. Darrell, "Conditional Random Fields for Object Recognition," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, 2005, pp. 1097–1104.
35. C. Yu, G. Liu, , and J. H. L. Hansen, "Acoustic feature transformation using ubm-based lda for speaker recognition," *in Proc. of Interspeech*, pp. 1851–1854, 2014.
36. X. Li and M. Akagi, "Multilingual Speech Emotion Recognition System based on a Three-layer Model," *in Proc. of Interspeech*, pp. 3606–3612, 2016.