Automatic Spoken Language Identification Using Emotional Speech

Panikos Heracleous[1], Akio Yoneyama[1], Kohichi Takai[1,2], and Keiji Yasuda[1,2]

[1] KDDI Research, Inc., Japan
{pa-heracleous,yoneyama}@kddi-research.jp
[2] Nara Institute of Science and Technology, Japan
takai.koichi.tc1@is.naist.jp,ke-yasuda@dsc.naist.jp

Abstract. Spoken language identification (LID) is the process of automatically recognizing the language from the uttered speech of an unknown speaker. Automatic recognition of language spoken is of vital importance in human-computer interaction and its applications. It can be applied in speech-to-speech translation systems, at call centers to re-route incoming calls to native speaker operators, and in speaker diarization in multilingual environments. The majority of studies which utilized LID systems focused solely on the use of neutral (i.e., normal) speech. However, in real applications and for comprehensive research investigations, the use of emotional speech in LID is crucial. The current study aims at investigating the effectiveness and performance of a deep neural networks (DNNs) based LID system when emotional speech is used.

Keywords: Spoken language identification · Emotional speech · Deep neural networks.

1 Introduction

Several studies have investigated spoken language identification. The approaches presented are categorized based on the features they employ. Language identification systems are categorized as the acoustic-phonetic approach, the phonotactic approach, the prosodic approach, and the lexical approach. In phonotactic systems [11], sequences of recognized phonemes obtained from phone recognizers are modeled. In acoustic modeling based systems, however, each recognized language is modeled by using different features. Although significant improvements in LID have been achieved using phonotactic-based approaches, most state-of-the-art systems rely on acoustic modeling [4, 10, 12, 9, 8, 16, 5, 14].

In the current study, a DNN-based [7] approach is introduced capable of classifying emotional speech produced in English, German, and Japanese languages. The system was evaluated and the results obtained were compared with those achieved when neutral speech was applied.

2 Methods

2.1 Speech Corpora

Three languages were considered namely, English, German, and Japanese. The English IEMOCAP database [3] is an acted, multimodal, and multi-speaker database collected at the SAIL lab of the University of Southern California and contains 12 hours of audiovisual data produced by ten actors. Specifically, the IEMOCAP database includes video, speech, motion capture of the face, and text transcriptions. The database consists of dyadic sessions where actors performed improvisations or scripted scenarios specifically selected to elicit emotional expression. The IEMOCAP database is annotated by multiple annotators into the several categorical labels of anger, happiness, sadness, neutrality, as well as the dimensional labels of valence, activation, and dominance. In the current study, categorical labels were used to classify the emotional states of neutral, happy, angry, and sad. For training, 1000 instances were used, and 200 instances were used for testing.

The German database used was the Berlin Emo-DB database [2], which includes seven emotional states: anger, boredom, disgust, anxiety, happiness, sadness, and neutral speech. For training, 280 instances were used, and 152 instances were used for testing.

Four professional female actors simulated Japanese emotional speech. These comprised neutral, happy, angry, and sad emotional states. Fifty-one utterances for each emotion was produced by each speaker. In total, 512 utterances were used for training, and 256 utterances were used for testing. The remaining utterances were excluded due to poor speech quality.

2.2 Feature extraction

In speaker recognition, Gaussian supervectors are widely used as features. The supervectors are constructed by concatenating the means of adapted Gaussian mixture models (GMMs). Although, significant improvements have been obtained using Gaussian supervectors, the main disadvantage of GMM supervectors is the high dimensionality, which imposes high computation and memory costs.

To overcome these problems, the i-vector paradigm [6] was introduced. The i-vectors represent the whole utterance with a small number of factors explaining the variability of speaker, channel, and language. An input utterance can be modeled as follows:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

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.

In automatic speech recognition, speaker recognition, emotion recognition, and language identification, mel-frequency cepstral coefficients (MFCCs) [13] are among the most popular and most widely used acoustic features. Therefore, this study similarly used 12 MFCCs concatenated with shifted delta cepstral (SDC) coefficients [1, 15] to form feature vectors of length 112 in modeling the languages and emotions being identified. 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 of dimension 100 used in emotion and spoken language identification modeling and classification.

2.3 Classification Methods

The classification experiments were based on DNNs. The DNN is an important method in machine learning that has been applied in many areas. A DNN is a feed-forward neural network with many (i.e., more than one) hidden layers. The main advantage of DNNs compared with shallow networks is the better feature expression and the ability to perform complex mapping. Deep learning explains several of the most recent breakthroughs in computer vision, speech recognition, and agents that achieved human-level performance in several games, such as Go and Poker. The DNN architecture used in the current experiment is a standard fully connected feedforward network with four hidden layers with 64 units followed by a Softmax layer for classification. All neurons employed the ReLU activation function, and 15% dropout was used to regularize the network. Stochastic Gradient Descent with Nestrov initialization and 0.9 momentum was employed for training (learningrate=0.01). Data were presented to the network in 500 epochs without early stopping.

3 Results

Table 1 shows the results obtained when using normal speech compared with using emotional speech. As is shown, when using normal speech a 97.5% average recall was achieved. This results is very promising and shows the effectiveness of using the proposed method for spoken language identification. Regarding the individual recalls, the English language shows perfect identification, with slightly lower recalls in the case of German and Japanese languages.

In the case of using emotional speech, the average recall was 93.8%. This rate is lower compared to normal speech. However, the recalls are still comparable and they show that no additional difficulties occurred in LID when using emotional speech. By performing the t-test, the two-tailed P value was 0.3410. By conventional criteria, this difference is considered to be not statistically significant.

4 Conclusions

The current study presented a method for automatic language identification using emotional speech. The results obtained were very promising and showed

Table 1. Spoken language identification recalls [%] using English, German, and Japanese speech data.

Speech data	Language			
	English	German	Japanese	Average
Normal	100.0	95.5	97.0	97.5
Emotional	97.4	87.6	96.5	93.8

that when using emotional speech, spoken language identification does not face any additional difficulties compared to normal speech. Specifically, for language identification using English, German, and Japanese speech, the average recall was 93.8%, slightly lower than the average recall when using normal speech. However, the recalls were closely comparable and the differences between normal and emotional recalls were not statistically significant. Currently, experiments using a larger number of languages are in progress.

References

1. Bielefeld, B.: Language identification using shifted delta cepstrum. In Fourteenth Annual Speech Research Symposium (1994)
2. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech. in Proc. of Interspeech pp. 1517–1520 (2005)
3. Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., Narayanan, S.: IEMOCAP: Interactive emotional dyadic motion capture database. Journal of Language Resources and Evaluation p. 335–359 (2008)
4. Cole, R., Inouye, J., Muthusamy, Y., Gopalakrishnan, M.: Language identification with neural networks: a feasibility study. in Proc. of IEEE Pacific Rim Conference p. 525–529 (1989)
5. Dehak, N., A.T.-Carrasquillo, P., Reynolds, D., Dehak, R.: Language recognition via ivectors and dimensionality reduction. in Proc. of Interspeech p. 857–860 (2011)
6. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing **19(4)**, 788–798 (2011)
7. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups ," IEEE Signal Processing Magazine **29, Issue:6**, 82–97 (2012)
8. Jiang, B., Song, Y., Wei, S., Liu, J.H., V.McLoughlin, I., , Dai, L.R.: Deep bottleneck features for spoken language identification. PLos ONE **9(7)**, 1–11 (2010)
9. L.-Moreno, I., G.-Dominguez, J., Plchot, O., Martinez, D., G.-Rodriguez, J., Moreno, P.: Automatic language identification using deep neural networks. in Proc. of ICASSP p. 5337–5341 (2014)
10. Leena, M., Rao, K.S., Yegnanarayana, B.: Neural network classifiers for language identification using phonotactic and prosodic features. in Proc. of Intelligent Sensing and Information Processing p. 404–408 (2005)

11. Li, H., Ma, B., Lee, K.A.: Spoken language recognition: From fundamentals to practice. in Proc. of the IEEE **101, no. 5**, pp. 1136–1159 (2013)
12. Montavon, G.: Deep learning for spoken language identification. in NIPS workshop on Deep Learning for Speech Recognition and Related Applications (2009)
13. Sahidullah, M., Saha, G.: Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition. Speech Communication **54 (4)**, 543–565 (2012). https://doi.org/doi:10.1016/j.specom.2011.11.004
14. Shen, P., Lu, X., Liu, L., Kawai, H.: Local fisher discriminant analysis for spoken language identification. in Proc. of ICASSP p. 5825–5829 (2016)
15. T.-Carrasquillo, P., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., Jr., J.D.: Approaches to language identification using gaussian mixture models and shifted delta cepstral features. in Proc. of ICSLP2002-INTERSPEECH2002 pp. 16–20 (2002)
16. Zazo, R., L.-Diez, A., G.-Dominguez, J., Toledano, D.T., , G.-Rodriguez, J.: Language identification in short utterances using long short-term memory (lstm) recurrent neural networks. PLos ONE **11(1): e0146917** (2016)