УДК 519.87

Speech-based Emotion Recognition and Speaker Identification: Static vs. Dynamic Mode of Speech Representation

Maxim Sidorov*
Wolfgang Minker†
Institute of Communications Engineering,
Ulm University,
Albert-Einstein-Allee, 43, Ulm, 89081
Germany
Eugene S. Semenkin‡
Informatics and Telecommunications Institute
Reshetnev Siberian State Aerospace University
Krasnoyarskiy Rabochiy, 31, Krasnoyarsk, 660037
Russia

Received 28.12.2015, received in revised form 24.02.2016, accepted 15.09.2016

In this paper we present the performance of different machine learning algorithms for the problems of speech-based Emotion Recognition (ER) and Speaker Identification (SI) in static and dynamic modes of speech signal representation. We have used a multi-corporal, multi-language approach in the study. 3 databases for the problem of SI and 4 databases for the ER task of 3 different languages (German, English and Japanese) have been used in our study to evaluate the models. More than 45 machine learning algorithms were applied to these tasks in both modes and the results alongside discussion are presented here.

Keywords: emotion recognition from speech, speaker identification from speech, machine learning algorithms, speaker adaptive emotion recognition from speech.
DOI: 10.17516/1997-1397-2016-9-4-518-523

Introduction

The main task of the speaker identification procedure is to determine who has produced a concrete utterance, but emotion recognition from speech can also reveal the emotional state of the person who has produced this utterance.

SI and ER can be used to improve a Spoken Dialogue System (SDS). Furthermore, specific information about a speaker can lead to higher ER accuracy.

Meanwhile there are many open questions in these fields. Among them the choice of the modelling algorithm and speech signal features could be mentioned. Moreover, state-of-the-art approaches include different schemes of speech signal representation: a static mode results in a single feature vector for each utterance, whereas in a dynamic regime the corresponding feature

* maxim.sidorov@uni-ulm.de
† wolfgang.minker@uni-ulm.de
‡eugenesemenkin@yandex.ru

set is calculated for each time window. The aim of the paper is to determine which combination of machine learning algorithms and types of speech signal representation should be used when the speech-based SI and ER problems are considered.

The rest of the paper is organized as follows. The databases used in the study are described in Section 1. Speech signal features are described in Section 2. Section 3 briefly describes the used machine learning algorithms. In section 4 we present the main results and draw some conclusions, and a description of directions for future work follows in Section 5.

1. Corpora Description

In this section we briefly describe the used corpora both for ER and Si tasks.

1.1. Emotion recognition databases

Emo-DB emotional database: The database [2] was recorded at the Technical University of Berlin and consists of labelled emotional German utterances which were spoken by 10 actors (5 females). Each utterance has one of the following emotional labels: neutral, anger, fear, joy, sadness, boredom and disgust.

LEGO emotional database: This database [3, 8, 9] comprises non-acted English (American) utterances which were extracted from the SDS-based bus-stop navigational system. The utterances are requests to the system spoken by real users of this system. Each utterance has one of the following emotional labels: angry, slightly angry, very angry, neutral, friendly and non-speech — critical noisy recordings or just silence.

UUDB: The database [7] consists of spontaneous Japanese speech through task-oriented dialogue, which was produced by 7 pairs of speakers (12 females), 4.737 utterances in total. Emotional labels for each utterance were created by 3 annotators on the 5-dimensional emotional basis (interest, credibility, dominance, arousal and pleasantness). To produce the labels for the classification task we have used only a pleasantness (or evaluation) and arousal axis. The corresponding quadrant (anticlockwise, starting in the positive quadrant, assuming arousal as abscissa) can also be assigned the emotional labels: happy-exciting, angry-anxious, sad-bored and relaxed-serene [10].

VAM-Audio database: The database [4] was created at Karlsruhe University and consists of utterances extracted from the popular German talk-show Vera-am-Mittag (Vera at the noon). The emotional labels of the first part of the corpus (speakers 1–19) were given by 17 human evaluators and the rest of the utterances (speakers 20–47) were labelled by 6 annotators on a 3-dimensional emotional basis (valence, activation and dominance). The emotional labelling was completed in a similar way to the UUDB corpora, using a valence (or evaluation) and arousal axis.

1.2. Speaker identification databases

RadioS database: Originally, it was a German radio talk-show, where people talk with a moderator about their private troubles. We have prepared a database based on the extraction of utterances from these talk-show recordings. The collection of data is still ongoing and by now it contains the utterances of 59 speakers.

PDA speech database: The recording of the corpus was performed at the Carnegie Mellon University using a PDA device. Each of 16 native speakers of American English reads about 50 sentences.

VAM-Video Dataset: This part of the VAM-Corpus [4] has no emotional labels but still can be used to evaluate speaker identification approaches. The number of speakers is 98.

The statistical description of the databases is in Tab. 1.

Table 1. Databases description

Database	Language	Full length(min.)	Num. of classes (Type)	File level duration		Class level duration	
				Mean(sec.)	Std. (sec.)	Mean(sec.)	Std. (sec.)
Berlin	German	24.7	7 (Emotions)	2.7	1.1	212.4	64.9
RadioS	German	235.6	59 (Speakers)	6.2	5.1	239.6	80.9
LEGO	English	133.0	6 (Emotions)	1.5	1.5	1330.2	1912.9
PDA	English	98.9	16(Speakers)	7.1	2.5	370.7	50.8
UUDB	Japanese	113.5	4 (Emotions)	1.4	1.7	1702.3	3219.8
VAM-A	German	47.8	4(Emotions)	3.0	2.2	717.2	726.4
VAM-V	German	75.8	98 (Speakers)	3.1	2.2	46.4	35.6

2. Speech signal features

The choice of the appropriate speech signal features for both problems is still an open question [11]. In this study we have chosen the most commonly used ones.

The following speech signal features were included into the feature vector: power, mean, root mean square, jitter, shimmer, 12 Mel-Frequency Cepstral Coefficients (MFCCs) and 5 formants. The mean, minimum, maximum, range and deviation of the following features were also used: pitch, intensity and harmonicity. While using the static mode, the averaged values of these features were applied, thus each speech signal is characterized by just one feature vector. On the other hand, the feature vectors have been extracted every 0.01 seconds due to the dynamic method of machine learning algorithm application.

The Praat system [1] was used for the feature extraction.

3. Machine learning algorithms

More than 45 algorithms were used for solving the problems. Whole databases were randomly divided into training and testing sets 70 vs. 30 (as percentages) correspondingly. Each algorithm can be assigned to one of the following classes.

Bayesian modelling. All these algorithms are based on the Bayesian theorem. The Simple Naive Bayes classifier and the Naive Bayes classifier with a kernel function as well as a Bayes Network were applied to the problems.

Artificial Neural Networks. This class of algorithms is based on the structural and functional modelling of the human brain. Such algorithms are capable of solving difficult modelling tasks. The state-of-the-art multi-layer perceptrons (MLP) and neural networks designed by evolutionary algorithms (Auto MLP) were applied to the classification tasks.

Support Vector Machine (SVM). SVM is the supervised learning algorithm based on the construction of a hyperplane or set of hyperplanes in a high- or infinite-dimensional space. These models can be used for classification, regression and other tasks. The SVM was applied to the tasks.

Decision tree based algorithms. This kind of model is based on a tree-like graph structure. The standard C4.5 and M5 algorithms for decision tree building, as well as the tree structure with logistic regression (LM Tree) and Naive Bayes classifiers (NB Tree) at its leaves were applied to the problems. A forest of random trees model and state-of-the-art tree structure used for the recursive partitioning (Decision Tree), as well as the tree model used for the information gain and variance and the pruning of its leaves using reduced-error pruning (REP Tree) were also applied to ER and SI problems.

Rule based algorithms. This type of algorithm involves growing rules corresponding to training data. The main advantage of such models is that they can be easily understood and represented in the first-order logic. The baseline RIPPER algorithm for growing rules, the hybrid algorithm for the decision table and naive Bayes classifier (DTNB) as well as the C4.5 and M5 rule-growing algorithms were applied to the problems. At each iteration the C4.5 and M5 algorithms build a decision tree model and make the best leaf into a rule.

Function fitting. This class of algorithms assumes that a model has some structure and the main task is to determine the appropriate parameters of this structure. The linear regression model assumes that a data set is linearly separable in the feature space. A multinomial logistic regression is also based on the logistic function and generalizes a simple logistic regression by allowing more than two discrete outcomes. These algorithms, as well as the Pace [12] regression were applied to the modelling. The PLS classifier is a wrapper classifier based on the PLS Filters which is able to perform predictions.

Lazy algorithms. This kind of algorithm uses only instances from a training set to create a class hypothesis of unknown instances. Basically they use different types of distance metrics between already known and unknown samples to produce a class hypothesis. The well-known K-Nearest Neighbours algorithm uses the Euclidian metric and the K-Star algorithm uses the Entropic-based metric.

Gaussian mixture models. This classifier builds the model of an object based on the weighted sum of Gaussian functions. To fit the parameters of such functions, the expectation maximization algorithm is used. The number of Gaussians in each model is equal to 32.

Fuzzy logic rules. The mixed fuzzy rule formation method was used to create a set of fuzzy logic rules.

4. Results of numerical evaluation

The algorithms were applied using the Weka [5] and RapidMiner [6] and KNIME programming systems. Some of them were implemented using the MATLAB programming system. Algorithms which have shown the highest values for accuracy during the static method of evaluations were combined in order to improve the performance of classification tasks. Thus, the best algorithms have solved the same problem, but the final decision about class hypothesis was made using the voting procedure.

The best algorithms for both the static and dynamic methods of evaluation are shown in Tab. 2. The artificial neural networks (MLP, AMLP and RBF) as well as the tree-based algorithms achieved the most precise emotional models during the static mode of application. As expected, the GMMs have built the best speaker identification models during the dynamic approach.

The algorithms which have achieved the highest accuracy level during the static evaluation method were included in the voting procedure. In this case a final decision is a hypothesis given

Table 2. Results of numerical evaluations on audio corpora

Database	Static (Alg.)	Dynamic (Alg.)	Vote
Berlin	0.71 (MLP)	0.7 (kNN)	**0.72**
LEGO	0.74 (AMLP)	0.68 (RBF)	**0.77**
UUDB	**0.92 (LM Tree)**	0.91 (AMLP)	0.91
VAM-A	0.7 (M5 Tree)	0.69 (MLP)	**0.74**
RadioS	0.89 (MLP)	**0.92 (GMMs)**	**0.92**
PDA	0.99 (LM Tree)	0.99 (GMMs)	**1**
VAM-V	0.54 (MLP)	0.6 (GMMs)	**0.66**

by the majority of algorithms. The proposed procedure has achieved the highest accuracy value for the most corpora, both the SI and ER tasks.

The static mode of speech data representation performed well for the ER task, whereas the dynamic one is the best choice for the SI task. Thus, the speaker-related information is uniformly distributed across the whole speech signal, whereas for the ER task, the most important features of the speech signal have a cumulative nature and could be captured by performing statistical analysis over the whole utterance.

5. Conclusions and future directions

It is evident that the classification accuracy depends strongly on the amount of speech data for each class. Therefore, the highest level of accuracy was achieved for the PDA corpus and the lowest one for the VAM-Video database (see the Number of classes and Class level length columns for these corpora in Tab. 1).

Such techniques as the Hidden Markov Model and neuro-fuzzy systems (such as ANFIS) are promising approaches for these types of classification tasks and the application of such algorithms is in our future plans.

Using the speaker identification procedure as an adaptation technique for emotion recognition could increase the robustness and accuracy of ER in the same way as for the problem of automatic speech recognition. The possibility of using such an adaptation technique is under examination now.

References

[1] P.Boersma, D.Weenink, Praat: doing phonetics by computer [Computer program]. Version 5.3.50, retrieved 21 May 2013 from http://www.praat.org/.

[2] F.Burkhardt et al., A database of German emotional speech, Proceedings of the Interspeech Conference, 2005, 1517–1520.

[3] M.Eskenazi, A.Black, A.Raux, B.Langner, Let's Go Lab: a platform for evaluation of spoken dialog systems with real world use, Proceedings of Interspeech Conference, Brisbane, Australia, 2008.

[4] M.Grimm, K.Kroschel, S.Narayanan, The Vera am Mittag German audio-visual emotional speech database, Multimedia and Expo, IEEE International Conference, 2008, 865–868.

[5] M.Hall et al., The WEKA Data Mining Software: An Update, *SIGKDD Explorations*, **11**(2009), no. 1, 10–18.

[6] I.Mierswa et al., YALE: Rapid Prototyping for Complex Data Mining Tasks, In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006.

[7] H.Mori et al., Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics, *Speech Communication*, **53**(2011), 36–50.

[8] A.Schmitt, B.Schatz, W.Minker, Modeling and predicting quality in spoken human-computer interaction, Proceedings of the SIGDIAL 2011 Conference, 2011, 173–184.

[9] A.Schmitt, T.Heinroth, J.Liscombe, On NoMatchs, NoInputs and BargeIns: Do Non-Acoustic Features Support Anger Detection? Proceedings of the SIGDIAL 2009. Conference, Association for Computational Linguistics, London, UK, 2009, 128–131.

[10] B.Schuller et al., Acoustic emotion recognition: A benchmark comparison of performances, In Automatic Speech Recognition and Understanding, IEEE Workshop, 2009, 552–557.

[11] M.Sidorov, A.Schmitt, S.Zablotskiy, W.Minker, Survey of Automated Speaker Identification Methods, Proceedings of Intellegent Environment 2013, Athens, Greece, 2013.

[12] Y.Wang, I.Witten, Modeling for optimal probability prediction, Proceedings of the Nineteenth International Conference in Machine Learning, Sydney, Australia, 2002, 650–657.

Распознавание эмоций и идентификация спикера по речевым сигналам: сравнение статического и динамического подходов к представлению речевых сигналов

Максим Сидоров
Вольфганг Минкер
Институт инженерных коммуникаций
Университет Ульма
Альберт-Эйнштейн-Аллея, 43, Ульм, 89081
Германия

Евгений С. Семенкин
Институт информатики и телекоммуникаций
Сибирский государственный аэрокосмический университет
Красноярский рабочий, 31, Красноярск, 660037
Россия

В статье рассматривается применение различных алгоритмов машинного обучения для задач распознавания эмоций и идентификации говорящего на основе речевых сигналов. Мы исследуем статический и динамический режимы представления речевого сигнала. Для проведения численных экспериментов и апробации рассмотренных подходов мы использовали 7 баз данных на немецком, английском и японском языках. Более 45 алгоритмов машинного обучения были применены для решения указанных задач в двух режимах представления речевого сигнала. В статье представлены результаты численных исследований и проведен их анализ.

Ключевые слова: распознавание эмоций и идентификация говорящего по речевым сигналам, алгоритмы машинного обучения, адаптивная процедура распознавания эмоций по речевым сигналам.