Knowledge-Based Systems xxx (xxxx) xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Posterior-thresholding feature extraction for paralinguistic speech classification☆

Gábor Gosztolya

MTA-SZTE Research Group on Artificial Intelligence, Szeged, 6720, Hungary

ARTICLE INFO

Article history:
Received 24 October 2018
Received in revised form 7 August 2019
Accepted 11 August 2019
Available online xxxx

Keywords:
Speech processing
Computational paralinguistics
Deep Neural Networks
Feature extraction
Classifier combination

ABSTRACT

The standard approach for handling computational paralinguistic speech tasks is to extract several thousand utterance-level features from the speech excerpts, and use machine learning methods such as Support Vector Machines and Deep Neural Networks (DNNs) for the actual classification task. In contrast, Automatic Speech Recognition handles the speech signal in small, equal-sized parts called *frames*. Although the speech community has developed techniques for efficient frame classification, these efforts have mostly been ignored in computational paralinguistics. In this study we propose a simple, three-step technique to utilize frame-level DNN training know-how in computational paralinguistics. We show that this method by itself provides good accuracy scores, and by combining it with the standard paralinguistic classification approach, we get close to the performance of heavyweight, state-of-the-art techniques such as Fisher vector analysis. However, our approach has the advantage that it can be easily realized by using standard speech recognition tools. To demonstrate the generic applicability of this three-step method proposed, we performed our experiments on four different corpora containing different paralinguistic tasks. Overall, we were able to achieve improvements over the baseline score in all four cases, leading to relative error reductions of up to 19%.

1. Introduction

Traditionally, the main focus of speech technology was Automatic Speech Recognition (ASR), where the task is to create the written transcription of an audio recording (an utterance) in an automatic way. Recently, however, extracting and identifying phenomena present in a speech signal other than the words uttered has gained further attention, and has formed a new area known as computational paralinguistics.

Of course, there are paralinguistic tasks which were dealt with even in the 90s such as identifying laughter events [1] and determining the speaker's emotional state [2,3]. Over the past decade, however, several other tasks have also gained attention such as conflict intensity estimation [4,5], detecting the amount of physical and cognitive load [6–9], detecting whether the speaker is drunk [10,11], and various medical applications like detecting Parkinson's disease, Alzheimer's disease and depression [12–15]. The standardization of methods, tools, evaluation metrics and the appearance of publicly available datasets was aided by the Interspeech Computational Paralinguistic Challenge (ComParE), held annually from 2009 [16].

ASR and computational paralinguistics are different by nature at two distinct levels. Firstly, ASR focuses on the words uttered, and treats everything else (the speaker's actual emotions, his physical load, blood alcohol level, etc.) as noise which is to be ignored. In paralinguistic tasks, however, we do not care about the individual words, but we are interested in other non-linguistic information present in human speech. The second difference is a more technical one: ASR handles the speech signal by dividing it into small, equal-sized excerpts called *frames*, on which local likelihoods are estimated, and these are then combined into a variable-length, utterance-level output (the phonetic or word-level transcription), usually via a Hidden Markov Model [17]. Therefore, when machine learning methods are used in ASR, they are usually applied at the frame level. In computational paralinguistics, however, each utterance is treated as one example, from which utterance-level features have to be extracted in some way. Classification also resembles general machine learning tasks instead of those common in ASR: there are only a few hundred examples instead of millions present in frame-level phoneme classification, hence researchers tend to prefer using Support Vector Machines (SVM, [18]) to Deep Neural Networks (DNNs, [19]).

In the standard solution for utterance-level feature extraction in computational paralinguistics, developed over the years mainly during the ComParE challenges (see e.g. [20–22]), first low-level descriptors such as energy, spectral, cepstral (MFCC) and voicing

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.104943.

https://doi.org/10.1016/j.knosys.2019.104943
0950-7051/

Table 1
The number of speakers and utterances in the training and test sets for the three databases used.

Dataset	Language	No. of classes	No. of utterances			Total duration (h:mm:ss)		
			Train	Test	Total	Train	Test	Total
Physical Load	German	2	770	319	1089	0:16:45	0:06:27	0:23:13
Emotion	Hungarian	4	831	280	1111	0:20:23	0:07:00	0:27:23
Eating Condition	German	7	945	469	1414	1:52:22	0:59:06	2:51:28
Cognitive Load (task *Reading sentence*)	English	3	1350	600	1950	1:28:21	0:40:58	2:09:19

related attributes are computed frame-wise; then, by using specific functionals like the mean and standard deviation, these are transformed into utterance-level features. Notice that no machine learning is performed at the frame level; however, in ASR (and in similar tasks such as laughter detection [23–25]) fine-tuned solutions exist on how frames should be classified. Unfortunately, these refined techniques are completely ignored in computational paralinguistics most of the time, and in the notable exceptions when they are not (e.g. [5,26,27]), machine learning is performed in a strongly task-dependent way.

Following our previous studies [28], we will present a method which synthesizes these two approaches. First, following standard ASR principles, we perform frame-level classification using Deep Neural Networks. Then, for the second step, we extract a new, utterance-level feature set from the frame-level DNN outputs; these new feature vectors are then utilized for utterance-level classification. We will show that this Posterior-Thresholding Feature Extraction (PTFE) approach is viable just on its own, as the results obtained this way appear to be close to those got using the standard paralinguistic approach. However, when we combine the predictions of the two techniques, we get accuracy scores that notably exceed those got by the baseline technique, falling closer to those obtained by heavyweight, state-of-the-art methods such as Fisher vector analysis of an Acoustic Background Model [26], which are not straightforward to utilize and are computationally very expensive. In contrast, our proposed method can be easily realized by relying on standard ASR tools, and there is no meta-parameter to be set.

Note that the proposed technique is somewhat similar to the Bag-of-Audio-Words (BoAW) representation [29], which has become popular recently [30,31]. In the BoAW approach, the frame-level feature vectors (e.g. MFCCs) of the training set are clustered to obtain the so-called *codebook* (the list of cluster centers). Each utterance is represented by the (normalized) histogram of the clusters of its frame-level feature vectors; classifier training and evaluation is realized by using these normalized histograms as utterance-level feature vectors. The key difference is that BoAW constructs the utterance-level feature representation in an unsupervised manner (i.e. via clustering), while within the PTFE method we employ machine learning for the same aim.

Another promising audio representation technique is wavelet-based multiresolution analysis [32]. The biggest advantage of multiresolution analysis is its denoising capability, since its focus lies in the area of noise-robust speech recognition [33–35]. In this study, however, we will focus on processing human speech recorded in (relatively) silent environments.

Lastly, we would also like to mention a further advantage of the proposed PTFE technique: since it contains no task-specific component, it can be expected to work task-independently. To demonstrate this, we validate the proposed workflow in four different paralinguistic tasks, ranging from estimating the level of physical load to emotion detection.

2. The datasets used

Next, we will describe the datasets we utilized in our experiments. To demonstrate the general utility of the proposed PTFE algorithm, we used four databases, which vary both in their topic and in their recording conditions (e.g. microphones, background noise, language). Table 1 contains some key properties of these paralinguistic datasets.

2.1. The Munich Biovoice corpus

The first dataset, called the Munich Biovoice Corpus [36], contains the utterances of 19 subjects (4 female and 15 male) of three nations (Chinese, German and Italian) speaking in German, after both light and heavy physical exercise. The subjects pronounced sustained vowels and read a short story, which was recorded by using two different microphones. (Besides the audio recordings, heart rate and skin conductivity were also monitored, but these measurements were not used in the current study.) This dataset was later used in the Interspeech ComParE 2014 Physical Load Sub-Challenge [21]; we will refer to it as the **Physical Load** dataset.

2.2. The Hungarian Emotion corpus

The Hungarian Emotion Database [37] contains sentences from 97 Hungarian speakers who participated in television programmes. A large portion of the segments were selected from spontaneous continuous speech rich in emotions (e.g. talk shows, reality shows), while the rest of the database came from improvised entertainment programmes. Note that, although actors tend to overemphasize emotions while acting, in improvisation their performance is more similar to real-life emotions [38].

Four emotion categories were defined: Anger, Joy, Neutral and Sadness. We defined our custom training and test sets, because in the original split these were not speaker-independent ones (as it was not a typical requirement at the time of recording). Our training set consisted of 831 segments, while the test set had 280 utterances. Due to this re-partitioning, our results presented here cannot be directly compared to those presented in the earlier studies (i.e. [37,38]), but classification accuracy scores around 66%–70% were reported. We will refer to this corpus as the **Emotion** dataset.

2.3. The iHEARu-EAT corpus

The *iHEARu-EAT* database [39] contains the utterances of 30 people recorded while they were speaking and eating at the same time. Six types of food were used along with the "no food" class, resulting in seven classes overall. For each speaker and food type, seven utterances were recorded; some subjects refused to eat certain types of foods, resulting in a total of 1414 utterances. Although this dataset can be used primarily to test machine learning and signal processing techniques, Hantke et al. also anticipated several possible future applications [39]. This dataset was used in the Interspeech ComParE 2015 Eating Condition Sub-Challenge [22]. Later on we will refer to this dataset as the **Eating Condition** dataset.

Please cite this article as: G. Gosztolya, Posterior-thresholding feature extraction for paralinguistic speech classification, Knowledge-Based Systems (2019) 104943, https://doi.org/10.1016/j.knosys.2019.104943.

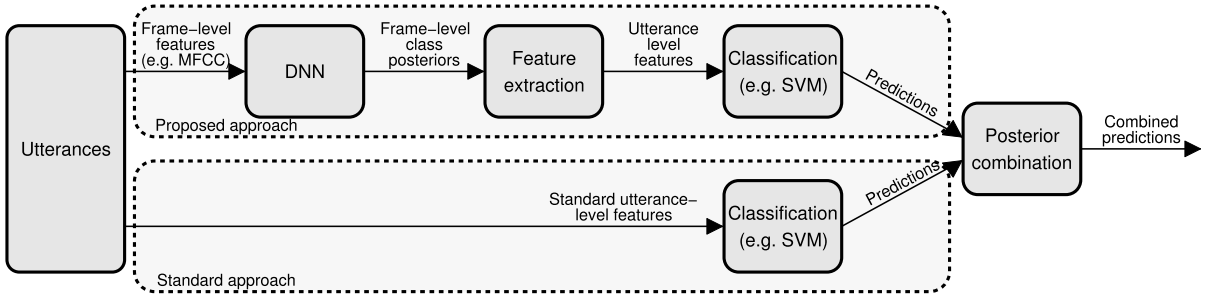

Fig. 1. The workflow of the proposed paralinguistic processing scheme.

2.4. The cognitive load with speech and EGG corpus

The purpose of the Cognitive Load with Speech and EGG database [40] is to evaluate algorithms which detect the cognitive load and working memory of speakers during speech. It contains the utterances of 26 native Australian English speakers (20 males and 6 females) performing 'span' tasks which require participants to recall a number of concepts or objects in the presence of distractors. The speakers had to perform three types of tasks. The first one (*reading sentence*) required them to read a series of short sentences, indicate whether each was true or false, and then remember a single letter presented briefly between sentences. Three different cognitive load levels were defined: low when remembering after one sentence, medium when remembering after two sentences, and high after the third, fourth and fifth sentences. The remaining two tasks were variants of the Stroop test [41]: the speakers had to name the font color of words corresponding to different color names. At the low level, the words and the colors were congruent, while at the medium and high levels they were not.

Since the three tasks performed were different by nature, it was advised that we train distinct classifier models for them (for the details, see [21]). However, due to the distribution of utterances, this results in fairly tiny datasets for the two Stroop tasks: from the 1674 utterances of the training set, 1350 belong to the reading span sentence task, while only 162–162 recordings contain speech recorded during the two Stroop test variations. Since the proposed PTFE algorithm relies on the frame-level DNN posteriors, and frame-level DNN training requires a relatively large database, we will just use the *reading sentence* task in our experiments.

This dataset was later used in the Interspeech ComParE 2014 Cognitive Load Sub-Challenge [21]; we will refer to it as the **Cognitive Load** dataset.

3. Posterior-thresholding feature extraction

Next, we will describe the proposed feature extraction and classification approach. (For the general scheme of the proposed workflow, see Fig. 1.) In the first step, we train a Deep Neural Network at the frame level. Then, in the second step, we calculate a new (utterance-level) feature vector based on the DNN outputs, which are used to train a Support Vector Machine to predict the actual paralinguistic phenomena. Lastly, we combine these predictions with those obtained using standard utterance-level features.

3.1. Frame-level classification

In the first step of our proposed workflow we train a Deep Neural Network with standard frame-level features (e.g. MFCC, PLP [17] and mel filter bank energies ("FBANK") [42]) as input,

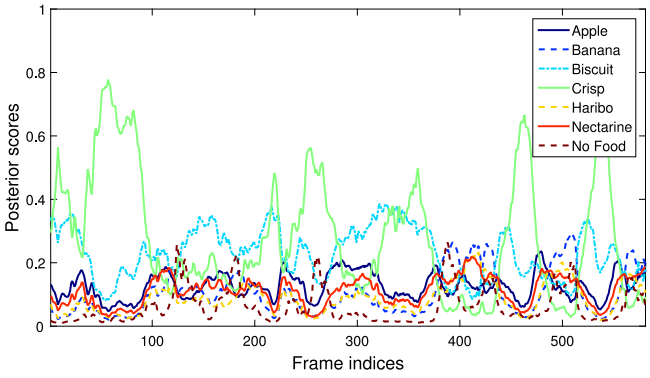

Fig. 2. The frame-level posterior scores obtained for an utterance of the Eating Condition corpus; the correct class for this example is 'Crisp'.

where the output neurons correspond to the actual, utterance-level class label for *each frame*. While doing so, of course, we do not expect that the frames will be classified very accurately. Since in most paralinguistic tasks the actual phenomena which distinguishes the different classes (e.g. breath intakes for physical load, coughs for cold, hesitation and silence in various kinds of dementia) is not present in every part of the utterance, it is impossible to detect it in a local manner (i.e. at the frame level). Still, since DNNs have proved to be quite robust in ASR, we may reasonably expect them to find the segments of the utterance which are specific to the given classes, and this will be reflected in the frame-level DNN outputs. Furthermore, as we do not intend to utilize the frame-level DNN outputs as-is, but we will process them further to obtain utterance-level feature vectors, it may be enough if they display some specific tendencies for utterances belonging to the given classes.

Fig. 2 shows a sample DNN output for the Eating Condition dataset. It is clear that the DNNs were able to find regions where sounds corresponding to the correct class (i.e. "Crisp" being eaten) are present; still, finding the correct class for the utterance may prove to be non-trivial.

3.2. Posterior-based feature extraction

In the second step of the proposed method, we extract features from the frame-level DNN outputs, which can then be used for *utterance-level* classification. Considering an actual application, this step could be generalized to be performed over some sliding window instead of the whole utterance. For most paralinguistic datasets, however, the manual annotation is given at the utterance level only, which does not allow continuous evaluation in our current study.

In the actual feature extraction step, we process the frame-level DNN outputs. The most straightforward solution is to classify each frame based on these likelihood scores, and obtain

Please cite this article as: G. Gosztolya, Posterior-thresholding feature extraction for paralinguistic speech classification, Knowledge-Based Systems (2019) 104943, https://doi.org/10.1016/j.knosys.2019.104943.

an utterance-level prediction by counting the ratio of frames classified as each possible class (*simple majority voting*, SMV [43]). For classification, the standard approach is to choose the class which has the largest posterior value for the given example (i.e. frame). For a binary case this is equivalent to thresholding the posterior estimates by the value 0.5. However, it is well known (see e.g. [19,44]) that the posterior estimates provided by a DNN contain valuable information, and this information might be exploited in some later step.

Because of this, we propose to employ *several* different threshold values. That is, using the step size parameter s, first we count the number of frames where the DNN output corresponding to the first class is greater than or equal to s, we normalize it by dividing it by the total number of frames in the utterance, and use this value as the first newly extracted feature. Next, we repeat this step using the values $2 \cdot s, 3 \cdot s, \ldots, 1$ as thresholds and for all the classes. Doing this for all the utterances, we extract a new feature set for all the utterances. For the pseudocode of this method, see Algorithm 1.

Note that, besides the fact that our approach can be viewed as a generalization of example-level classification (i.e. thresholding poster estimates using the value 0.5), it has another justification. Extracting the PT feature set is equivalent to calculating the *cumulative histogram* [45] of the frame-level posterior estimates. Cumulative histograms were previously used in ASR [46] as well as in several other tasks such as texture classification [47], handwritten character recognition [48] and analog-to-digital converter testing [49].

Algorithm 1 Posterior-Thresholding Feature Extraction

Require: *likelihoods*: the DNN outputs for the utterance
Require: f: the number of frames in the utterance
Require: N: the number of classes
Require: s: the step size ($s < 1$)
  $m := \lfloor 1/s \rfloor$
  for i := $1 \rightarrow N$ **do**
    for j := $1 \rightarrow m$ **do**
      $cnt := 0$
      for k := $1 \rightarrow f$ **do**
        if $likelihoods(k, i) \geq j \cdot s$ **then**
          $cnt := cnt + 1$
        end if
      end for
      $features((i-1) \cdot m + j) := cnt/f$
    end for
  end for
  return *features*

3.3. Utterance-level classification

This step is actually quite simple: using the feature vectors extracted in the previous step, we perform the classification of the utterances. Since our method is intended for general computational paralinguistic tasks, we suggest using Support Vector Machines as the classifier method, which has proved to be the most robust and efficient in computational paralinguistic tasks. Of course, some studies (see e.g. [50–52]) report that by carefully setting the meta-parameters, Deep Neural Networks can provide a similar performance. For the sake of simplicity, however, in this study we will utilize only SVMs for utterance-level classification.

3.4. Feature set combination

Although using the thresholded posteriors may prove to be beneficial for classification, we should not discard all other kinds of features, especially since the standard 'ComParE functionals' feature set proposed by Schuller et al. [20] has proved to be quite effective over the years on several different tasks (see e.g. [21,53,54]). Optimality is probably achieved via some combination of the proposed approach with this standard one. In our experience, as employing *late fusion* by averaging out the posterior estimates of the different approaches (e.g. feature sets) is a simple-yet-effective approach [28,55], we will apply this solution in our experiments.

4. Experimental setup

4.1. DNN parameters

At the frame level we trained a Deep Neural Network with 3 hidden layers, each containing 256 rectified neurons [56], and we used the softmax activation function in the output layer. We used our custom implementation for Nvidia GPUs, which achieved very good accuracy scores on several tasks and datasets (e.g. [57,58]). Training was performed on a 15-frame long sliding window, which is also a standard technique in phoneme classification within ASR [19,57].

We started with a fixed learning rate of 0.001 for the Physical Load and the Emotion corpora, and a 0.0001 for the Eating Condition and the Cognitive Load datasets. These values were determined by preliminary tests; the difference in the initial learning rates could probably be explained by the amount of training data required: the networks needed a larger initial learning rate when there was a smaller number of training examples available. During training, the learning rate was set via the *newbob* learn rate scheduler method [59]. In it, the (here frame-level) accuracy of the network is measured after each training iteration on a hold-out set, and we cease training when the accuracy score does not improve within a certain number of consecutive iterations. We used a random 10% of the training data for such a hold-out set; when the measured accuracy score did not improve for two consecutive epochs, we halted DNN training.

Note that we did not fine-tune the meta-parameters of frame-level DNN training (the number of hidden layers, sliding window width, etc.) at all for two reasons. Firstly, we assumed that the tendency of the posterior values can be exploited in the second step even if these posterior estimates are somewhat suboptimal. Secondly, in this study our aim was to present a general procedure which performs well as-is, without the need for carefully setting a number of meta-parameter values.

4.2. Frame-level feature sets

In this study we tested two types of frame-level feature vectors. The first one was the Mel-Frequency Cepstral Coefficients (MFCC, [17]), being quite popular in phoneme classification; we used 12 MFCC values along with energy, and their first and second order derivatives ("MFCC $+\Delta$ $+\Delta\Delta$"), which resulted in 39 attributes overall. As DNNs were shown to perform better on more primitive feature sets, we also experimented with relying on raw mel filter bank energies ("FBANK") of 40 bands; this feature set, after including the energy of the speech signal and the Δ and $\Delta\Delta$ values here as well, resulted in 123 feature values for each frame.

4.3. Feature extraction and utterance classification

In the subsequent feature extraction step, we used a step size of $s = 0.02$ for the thresholds, resulting in 50 features for each class; this meant feature sets with a size of 100, 200, 350 and 150

Please cite this article as: G. Gosztolya, Posterior-thresholding feature extraction for paralinguistic speech classification, Knowledge-Based Systems (2019) 104943, https://doi.org/10.1016/j.knosys.2019.104943.

for the Physical Load, Emotion, Eating Condition and Cognitive Load datasets, respectively. We applied the ν-SVM method with linear kernel using the LibSVM implementation [60]; the value of C was tested in the range $10^{\{-5,\ldots,2\}}$, just like in our previous paralinguistic studies (e.g. [28,52,55,61]).

Training was done in the way which is common in computational paralinguistics : the training meta-parameters (e.g. C for SVM) were determined in speaker-wise cross-validation (CV). To make the predictions for the test set, we trained an SVM model using all the training examples. Note that the Physical Load and Cognitive Load datasets were introduced in the ComParE 2014 Challenge [21] with separate training and development sets. However, for datasets with similar sizes, using speaker-wise cross-validation is viewed as standard practice in computational paralinguistics nowadays.

Due to the large number of speakers present in the Emotion dataset, we regarded speaker-wise cross-validation as unfeasible and used ten-fold cross-validation instead. Of course, all the utterances of a speaker were assigned to the same fold. Furthermore, since the class distribution for this corpus was highly imbalanced (which is typical for emotion datasets; see e.g. [62,63]), we upsampled the rarer classes.

4.4. Evaluation metrics

The accuracy of classification was measured via the Unweighted Average Recall (UAR) metric, it being the mean of the class-wise recall scores; this is the de facto standard evaluation metric on three datasets [22,39], and it is widely used in computational paralinguistics in general. We also report standard classification accuracy scores for each case; although, as for the Eating Condition and Cognitive Load datasets the class distribution is quite balanced, the UAR scores appear to be very similar to the corresponding classification accuracy scores.

4.5. Other paralinguistic approaches

We also classified the utterances following the standard paralinguistic approach, i.e. employing the 6373-sized 'ComParE functionals' feature set proposed by Schuller et al. [20], extracted by the `openSMILE` tool [64]. This set includes energy, spectral, cepstral (MFCC) and voicing-related frame-level attributes, from which specific functionals (like the mean, standard deviation, percentiles and peak statistics) are computed to provide utterance-level feature values.

For comparison, we also tested the Bag-of-Audio-Words (BoAW) representation [29]. We used the openXBOW package [65]; the 39-sized MFCC vectors were used as the frame-level inputs. Unfortunately, unlike PTFE, BoAW has a meta-parameter to be set: the codebook size (i.e. number of clusters). Since the size of PT feature sets ranged from 100 to 350, we tested codebook sizes of 32, 64, 128, 256, 512 and 1024 in the case of BoAW representation to roughly match feature set sizes.

The last approach we tested for reference was to combine the frame-level DNN outputs in other ways: we took their mean for each class within the given utterance, we combined them by multiplication, and we experimented with choosing the most probable class for each frame and then using simple majority voting of the frame-level class label hypotheses.

4.6. Feature standardization

Support Vector Machines, like most machine learning methods, are sensitive to the scale of the different features, hence they require feature standardization or normalization. However, for specific computational paralinguistic tasks it was shown (see e.g. [22,26,66,67]) that speaker-wise feature standardization might assist the subsequent classification steps. Therefore, in our preliminary tests we also experimented with this standardization approach for both the original and the PTFE feature sets. We made use of the annotated speaker IDs for the training set, while the speakers of the test set were determined by single Gaussian-based bottom-up Hierarchical Agglomerative Clustering with Generalized Likelihood Ratio (GLR) as the distance measure [26,68,69]. We found that speaker-wise standardization is useful both for the ComParE and the PTFE feature sets for the Eating Condition and the Cognitive Load datasets, so we standardized both feature sets this way before training and evaluating our SVM models. For the Physical Load and Emotion databases, however, we performed global standardization for all feature sets tested.

4.7. Prediction combination

As we mentioned in Section 3.4, we also experimented with combining the PT and the ComParE feature sets by taking the weighted mean of the utterance-level posterior estimates. The weights summed up to one, and we used a step size of 0.05. Optimal weights were set via cross-validation. For comparison, we also combined the MFCC BoAW features with the ComParE feature set in the same way.

4.8. Training set division

Notice that the proposed Posterior-Thresholding Feature Extraction workflow performs classifier training at two distinct levels: first it trains a DNN at the frame level, then it trains some other classifier (e.g. an SVM) at the utterance level, using the features extracted from the frame-level DNN outputs. However, DNNs are known to have a tendency of overfitting (see e.g. [70]), i.e. the posterior scores of the examples on which DNN training was performed will be biased towards the correct class. Extracting the PTFE features from the DNN training set and incorporating them into the SVM training set would create a mismatch among the (utterance-level) training and test sets, and probably harm classification accuracy. To avoid this, it would be beneficial to separate a subset of the training set for DNN training, and exclude these utterances from the utterance-level model training step. In practice, however, computational paralinguistic datasets tend to be of a fairly limited size, so discarding a part of training utterances might lead to a significant loss in model accuracy.

Because of this, we decided to split our training sets into two, equal-sized parts. We trained our frame-level DNN models on the first one, evaluated them on the second half of the training set and on the test set, and extracted the PTFE features. We trained SVM models on the second half of the training set in speaker-wise cross-validation mode, determined optimal SVM complexity, and evaluated these models on the test set. Next, we switched the roles of the two halves of the training set, and repeated the whole process. To combine the predictions of the SVM models on the test set, we averaged out their posteriors in an unweighted manner. This way we ensured that SVM models were trained on unbiased PTFE features, while all training examples were made use of.

5. Results

Tables 2 to 5 show the results obtained for the four datasets. Notice that using linear SVM with the ComParE feature set (used as our baseline scores) and the actual baseline values of the corresponding ComParE challenges differ to a certain extent. This is due to a number of reasons. Firstly, we used the `libSVM` [60] Support Vector Machine implementation instead of Weka, being

Please cite this article as: G. Gosztolya, Posterior-thresholding feature extraction for paralinguistic speech classification, Knowledge-Based Systems (2019) 104943, https://doi.org/10.1016/j.knosys.2019.104943.

Table 2
The results obtained on the Physical Load corpus.

Approach		Cross-validation		Test	
		Acc.	UAR	Acc.	UAR
ComParE functionals feature set (baseline)		68.1%	67.9%	70.8%	71.0%
MFCC BoAW features		65.7%	65.7%	73.4%	73.2%
ComPare + MFCC BoAW features		69.5%	69.3%	73.0%	73.2%
MFCC	Frame-level DNN outputs (mean)	65.9%	65.9%	55.8%	54.8%
	Frame-level DNN outputs (product)	66.7%	66.7%	54.9%	53.8%
	Frame-level DNN outputs (majority voting)	66.2%	66.1%	54.5%	53.5%
	Posterior-thresholding (PT) features	66.5%	66.3%	69.0%	69.0%
	ComParE + PT features	70.5%	70.3%	74.0%	74.0%
FBANK	Frame-level DNN outputs (mean)	65.0%	64.7%	53.3%	53.8%
	Frame-level DNN outputs (product)	59.8%	59.1%	53.0%	53.7%
	Frame-level DNN outputs (majority voting)	65.2%	64.9%	52.4%	52.9%
	Posterior-thresholding (PT) features	64.0%	63.7%	73.0%	73.1%
	ComParE + PT features	68.8%	68.6%	73.4%	73.5%
ComParE 2014 baseline (Schuller et al., [21])		–	–	–	71.9%
Chance		50.0%	50.0%	50.0%	50.0%

Table 3
The results obtained on the Emotion corpus.

Approach		Cross-validation		Test	
		Acc.	UAR	Acc.	UAR
ComParE functionals feature set (baseline)		67.6%	54.5%	76.8%	60.3%
MFCC BoAW features		66.8%	55.4%	81.1%	56.2%
ComPare + MFCC BoAW features		67.3%	56.2%	81.4%	57.5%
MFCC	Frame-level DNN outputs (mean)	65.5%	46.6%	84.3%	51.4%
	Frame-level DNN outputs (product)	65.0%	47.3%	84.3%	53.2%
	Frame-level DNN outputs (majority voting)	65.6%	47.2%	84.6%	51.5%
	Posterior-thresholding (PT) features	60.8%	54.7%	75.4%	58.2%
	ComParE + PT features	68.2%	56.2%	78.9%	61.0%
FBANK	Frame-level DNN outputs (mean)	67.9%	48.8%	81.8%	50.3%
	Frame-level DNN outputs (product)	68.2%	50.2%	81.1%	51.3%
	Frame-level DNN outputs (majority voting)	67.9%	48.8%	82.1%	51.6%
	Posterior-thresholding (PT) features	67.3%	49.5%	75.0%	54.7%
	ComParE + PT features	70.8%	56.4%	78.9%	61.3%
Chance		25.0%	25.0%	25.0%	25.0%

the standard in the ComParE Challenges. Secondly, we relied on speaker-wise cross-validation instead of using a separate development set, which might slightly affect the meta-parameter setting. The third, and perhaps most important cause of the differences is the speaker-wise standardization technique applied for the Eating Condition and Cognitive Load datasets, as official challenge baselines were obtained via global standardization. Of course, we will treat the former scores as our baselines, since they mirror our experimental protocol.

Table 2 contains the results obtained for the **Physical Load** corpus. We can see that using the BoAW representation alone yielded scores which almost matched those of the ComParE functionals feature set in the cross-validation set-up, while it outperformed the baseline on the test set. Combining these two approaches, however, yielded no further improvement. Fusing the frame-level DNN outputs into utterance-level hypotheses by mean, product and majority voting led to nice UAR values in CV, as the resulting scores were around 66% for both frame-level feature sets tested (i.e. MFCC and FBANK). However, on the test set these strategies performed quite poorly: the UAR scores of 52.9–54.8% only slightly exceed the 50% achievable by random guessing in this binary problem. Compared to these values, using only the Posterior-Thresholding (PT) features and training an SVM led to much higher values, although the scores obtained for the MFCC case are slightly below the baseline. In our opinion this indicates that the paralinguistic phenomena appearing in this dataset can be captured locally, in a frame-based manner via DNNs, but only to a limited extent, and combining the DNN outputs at the utterance level is not trivial. From this aspect, extracting the Posterior-Thresholding features and training another classifier model (this time an SVM) appears to be a much better strategy than the trivial combination approaches of posterior mean, posterior product and simple majority voting.

We obtained the highest accuracy and UAR scores when we utilized both the standard ComParE feature set and the Posterior-Thresholding features: combining the predictions got by using the ComParE and the PT (MFCC) feature sets via late fusion led to the highest score (74.0%), meaning a relative error reduction (RER) score of over 10%. When we relied on the FBANK features, we got a slightly lower improvement: the UAR score of 73.5% corresponds to an RER value of about 8%.

For the **Emotion** dataset (see Table 3), the BoAW features led to slightly lower scores than our baseline, and the combination of the two approaches could not exceed the baseline either. Combining the frame-level DNN outputs by mean, product and majority voting led to UAR values around 47%–50% (CV) and about 50%–53% (test); notice, however, that the corresponding classification accuracy scores are much higher. The reason for this is probably that the other classification approaches all employed upsampling, leading to a more balanced performance class-wise. However,

Please cite this article as: G. Gosztolya, Posterior-thresholding feature extraction for paralinguistic speech classification, Knowledge-Based Systems (2019) 104943, https://doi.org/10.1016/j.knosys.2019.104943.

Table 4
The results obtained on the Eating Condition corpus.

Approach		Cross-validation		Test	
		Acc.	UAR	Acc.	UAR
ComParE functionals feature set (baseline)		74.5%	74.3%	75.3%	74.8%
MFCC BoAW features		66.0%	65.8%	70.1%	69.6%
ComPare + MFCC BoAW features		75.8%	75.6%	75.7%	75.2%
MFCC	Frame-level DNN outputs (mean)	41.5%	40.9%	43.7%	43.1%
	Frame-level DNN outputs (product)	41.6%	41.1%	44.1%	44.0%
	Frame-level DNN outputs (majority voting)	41.6%	41.1%	45.4%	44.7%
	Posterior-thresholding (PT) features	64.6%	62.1%	66.3%	65.8%
	ComParE + PT features	76.1%	75.9%	78.9%	78.6%
FBANK	Frame-level DNN outputs (mean)	36.3%	35.4%	42.2%	41.1%
	Frame-level DNN outputs (product)	36.4%	35.5%	41.6%	40.6%
	Frame-level DNN outputs (majority voting)	35.3%	34.5%	41.8%	40.6%
	Posterior-thresholding (PT) features	57.6%	56.9%	64.2%	63.3%
	ComParE + PT features	75.8%	75.5%	78.5%	78.0%
ComParE 2015 baseline [22]		–	61.3%	–	65.9%
Chance		14.3%	14.3%	14.3%	14.3%

Table 5
The results obtained on the reading sentence task of the Cognitive Load corpus.

Approach		Cross-validation		Test	
		Acc.	UAR	Acc.	UAR
ComParE functionals feature set (baseline)		64.8%	62.9%	64.8%	63.4%
MFCC BoAW features		56.8%	54.0%	55.3%	51.9%
ComPare + MFCC BoAW features		64.8%	62.9%	64.8%	63.4%
MFCC	Frame-level DNN outputs (mean)	44.0%	33.3%	44.0%	33.3%
	Frame-level DNN outputs (product)	44.0%	33.3%	44.0%	33.3%
	Frame-level DNN outputs (majority voting)	44.0%	33.3%	44.0%	33.3%
	Posterior-thresholding (PT) features	49.5%	44.2%	54.3%	48.5%
	ComParE + PT features	64.8%	62.9%	64.8%	63.4%
FBANK	Frame-level DNN outputs (mean)	44.0%	33.3%	44.0%	33.3%
	Frame-level DNN outputs (product)	44.0%	33.3%	44.0%	33.3%
	Frame-level DNN outputs (majority voting)	44.0%	33.3%	44.0%	33.3%
	Posterior-thresholding (PT) features	54.1%	49.3%	56.7%	51.6%
	ComParE + PT features	64.8%	62.9%	64.8%	63.4%
ComParE 2014 baseline [21]		–	61.3%	–	61.5%
Chance		33.3%	33.3%	33.3%	33.3%

combining the frame-level posteriors favored the more frequent classes, resulting in higher classification accuracy and lower UAR values. Relying on the PT features, we slightly outperformed the BoAW approach, but the results remained below the baseline on the test set for both frame-level feature sets; but using late fusion to merge the predictions of the PT and the ComParE functionals approaches led to a slight increase in the UAR score on the test set (relative error reduction scores of roughly 3% and 4%, MFCC and FBANK case, respectively).

Our experiences on the **Eating Condition** dataset (see Table 4) were similar to those obtained on the previous two corpora. Combining the Bag-of-Audio-Words technique with the standard ComParE feature set brought a small improvement (0.4% absolute) on the test set, but this is probably not statistically significant. The PT feature set performed below the two standard approaches tested (i.e. ComParE and BoAW); however, it could be combined with the ComParE feature set quite efficiently, leading to RER scores of 13%–15%.

For the **Cognitive Load** dataset (see Table 5), the BoAW features were not useful at all: the 51.9% UAR score on the test set falls below the baseline score of 63.4%, and the combination via late fusion did not bring any improvement over the baseline either. We can see that this was so for the PT feature set as well, regardless of whether we relied on the MFCC or on the FBANK frame-level feature vectors; this is probably due to the low quality of DNN outputs for this task, which is also reflected in the UAR scores of 33.3% got by recombining the DNN outputs either via mean, product or simple majority voting.

Overall, the scores obtained by combining the ComParE functionals features with the PT ones always exceeded the baseline values, the only exception being the Cognitive Load corpus, where the scores remained unaffected. Examining the optimal late fusion weight values of the ComParE and the PT features, the two types of features had roughly the same importance: the predictions obtained using the baseline feature set had a weight of 0.5 and 0.6, Physical Load and Eating Condition, respectively. On the Cognitive Load corpus, however, the ComParE feature set had an optimal weight of 1.0, indicating that the Posterior-Thresholding features were not useful at all with late fusion.

Overall, it can be seen that using the Posterior-Thresholding features and training an SVM model at the utterance level led to UAR values much higher than those which can be obtained by random guessing, for all four datasets. The PT features also had the advantage that they describe the phenomena actually present in the utterances in a completely different way than the ComParE functionals or the BoAW features do, allowing an efficient combination that managed to improve both the traditional classification accuracy and the UAR scores in three of the four cases. Next, we

Please cite this article as: G. Gosztolya, Posterior-thresholding feature extraction for paralinguistic speech classification, Knowledge-Based Systems (2019) 104943, https://doi.org/10.1016/j.knosys.2019.104943.

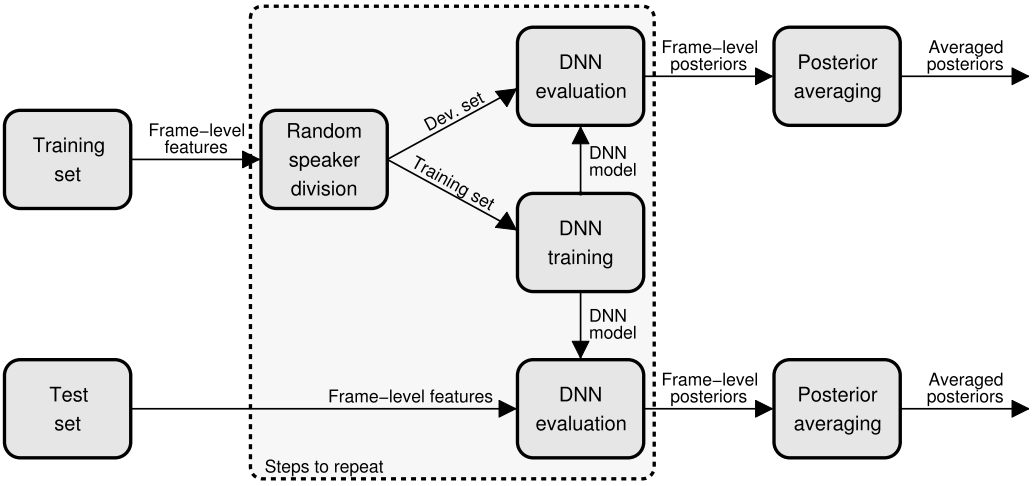

Fig. 3. The workflow of the proposed method for obtaining frame-level posterior estimates.

will present an approach that leads to even better frame-level posterior estimates.

6. Obtaining robust frame posteriors

Our results indicated that the Posterior-Thresholded features can be utilized to classify whole utterances despite the fact that the frame-level posterior estimates are local by nature. However, it was not straightforward to produce frame-level DNN outputs without a bias towards the correct class, while also making use of all training examples in the second, utterance-level training step. The approach proposed in Section 4.8 (dividing the training set into two parts, one used for DNN training and one used for SVM training, then switching their roles and repeating these steps) managed to satisfy both requirements; still, since we trained two frame-level DNNs on two different subsets of the utterances, the PT features could be expected to differ significantly on the two subsets of the training set, making the process of combining their predictions on the test set quite complicated. Next we will introduce a simple procedure for getting robust frame-level posterior estimates for all the utterances.

6.1. The proposed workflow

The workflow we propose for obtaining robust frame-level posterior estimates is an iterative process. For each step, we randomly split the training set into two distinct subsets: the utterances of half the speakers form the group "DNN Training Set", while the remaining half form the "DNN Development Set". We train a DNN on the utterances belonging to the first group, and evaluate this neural network on the frames of the second group (i.e. the actual development set) and of the test set. Repeating these steps several times and averaging out the resulting frame-level DNN outputs, we get posteriors estimates which can be expected to be quite robust. (See Fig. 3.)

In our experiments we used 250 iterations; this way, the frame-level posterior estimates were derived from about 125 models on the (full) training set, and from all the 250 models on the test set. We judged 250 iterations to be sufficient to provide robust estimates of the frame-level DNN outputs, while also keeping the time requirement of the process at an acceptable level.

6.2. Results

Note that in this step we did not test both frame-level feature sets any more for each dataset: since for the Physical Load and the Eating Condition corpora, the UAR values obtained by relying on MFCCs always exceeded those got via FBANK, we will only present the scores got via MFCCs. For the Emotion and the Cognitive Load corpora, however, we will again test both approaches (see Table 7).

Tables 6 to 9 show the results obtained for the four datasets used in this study. Since we changed only the way the frame-level posterior estimates were calculated, the accuracy and UAR values achieved by the ComParE functionals feature set remained unchanged. The utterance-level scores got by aggregating the frame-level posterior estimates via mean, product and simple majority voting, however, rose significantly for the test set of two datasets: in the case of the Physical Load corpus, they increased from about 54% to 62%, while for the Eating Condition corpus they rose from about 44% to 45%–50%. This, in our opinion, reflects the higher quality of frame-level DNN outputs: since these posterior estimates now come from several DNN models, they are less prone to noise introduced by random DNN weight initialization and random training instance selection.

For the Emotion dataset, we can see that the UAR scores on the test set remained around 47%, but in the CV set-up they fell from 57% to 47%. In our opinion this also demonstrates the improved robustness of the DNN posterior estimates, as now the UAR scores show a similar tendency for the two subsets, which can also be expected to improve classifier fusion quality. As regards the Cognitive Load corpus, all the scores remained at 33.3%, since the DNN outputs corresponding to the most common class (L1) were the highest ones for practically all the frames.

When we trained Support Vector Machines using the Posterior-Thresholding features, we observed a slight increase in the UAR score for the test set of the Physical Load and the Emotion corpora (improvements below 1% absolute), but for the remaining two datasets the increase was significant: the UAR scores rose from 65.8% to 69.8%, and from 48.5% to 56.5% for the Eating Condition and Cognitive Load corpus, respectively. When we combined the PTFE feature set with the baseline ComParE functionals one via late fusion, we also observed a significant increase in the accuracy and UAR scores. The only exception was again the Cognitive Load corpus, where the UAR scores obtained via combination managed to outperform the baseline score of 63.4%, but only by 0.5 — 0.8% absolute.

Please cite this article as: G. Gosztolya, Posterior-thresholding feature extraction for paralinguistic speech classification, Knowledge-Based Systems (2019) 104943, https://doi.org/10.1016/j.knosys.2019.104943.

Table 6
The results obtained on the Physical Load corpus.

Approach		Cross-validation		Test	
		Acc.	UAR	Acc.	UAR
ComParE functionals feature set (baseline)		68.1%	67.9%	70.8%	71.0%
MFCC	Frame-level DNN outputs (mean)	67.6%	67.3%	62.7%	62.1%
	Frame-level DNN outputs (product)	67.5%	67.1%	62.7%	62.2%
	Frame-level DNN outputs (majority voting)	68.3%	67.9%	63.0%	62.5%
	Posterior-thresholding (PT) features	67.2%	66.9%	69.3%	69.2%
	ComParE + PT features	70.5%	70.3%	74.6%	74.7%
ComParE 2014 baseline [22]		–	–	–	71.9%
Prosodic and ASR-derived features [71]		–	71.8%	–	73.9%
MRMR Filter feature selection [9]		–	–	–	75.4%

Table 7
The results obtained on the Emotion corpus.

Approach		Cross-validation		Test	
		Acc.	UAR	Acc.	UAR
ComParE functionals feature set (baseline)		67.6%	54.5%	76.8%	60.3%
MFCC	Frame-level DNN outputs (mean)	72.2%	47.3%	82.5%	45.4%
	Frame-level DNN outputs (product)	72.2%	47.7%	82.9%	47.7%
	Frame-level DNN outputs (majority voting)	72.3%	47.3%	82.1%	45.1%
	Posterior-thresholding (PT) features	61.7%	50.3%	76.4%	57.1%
	ComParE + PT features	67.4%	54.7%	77.9%	61.9%
FBANK	Frame-level DNN outputs (mean)	72.2%	45.8%	81.1%	46.3%
	Frame-level DNN outputs (product)	72.4%	45.9%	81.4%	47.6%
	Frame-level DNN outputs (majority voting)	72.6%	46.5%	80.7%	45.9%
	Posterior-thresholding (PT) features	65.3%	55.9%	70.4%	55.4%
	ComParE + PT features	70.9%	58.0%	79.6%	64.4%

Table 8
The results obtained on the Eating Condition corpus.

Approach		Cross-validation		Test	
		Acc.	UAR	Acc.	UAR
ComParE functionals feature set (baseline)		74.5%	74.3%	75.3%	74.8%
MFCC	Frame-level DNN outputs (mean)	46.2%	45.5%	48.0%	46.8%
	Frame-level DNN outputs (product)	47.6%	46.9%	50.7%	49.9%
	Frame-level DNN outputs (majority voting)	45.1%	44.4%	46.9%	45.6%
	Posterior-thresholding (PT) features	66.7%	66.2%	70.6%	70.0%
	ComParE + PT features	77.1%	76.9%	80.2%	79.7%
ComParE 2015 baseline [22]		–	61.3%	–	65.9%
Fisher Vector analysis with Acoustic Background Model [26]		–	78.9%	–	81.6%
Best result reported [26]		–	–	–	83.1%

Table 9
The results obtained on the reading sentence task of the Cognitive Load corpus.

Approach		Cross-validation		Test	
		Acc.	UAR	Acc.	UAR
ComParE functionals feature set (baseline)		64.8%	62.9%	64.8%	63.4%
MFCC	Frame-level DNN outputs (mean)	44.0%	33.3%	44.0%	33.3%
	Frame-level DNN outputs (product)	44.0%	33.3%	44.0%	33.3%
	Frame-level DNN outputs (majority voting)	44.0%	33.3%	44.0%	33.3%
	Posterior-thresholding (PT) features	56.8%	53.5%	59.5%	56.5%
	ComParE + PT features	65.0%	63.0%	65.8%	64.0%
FBANK	Frame-level DNN outputs (mean)	44.0%	33.3%	44.0%	33.3%
	Frame-level DNN outputs (product)	44.0%	33.3%	44.0%	33.3%
	Frame-level DNN outputs (majority voting)	44.0%	33.3%	44.0%	33.3%
	Posterior-thresholding (PT) features	57.9%	54.5%	57.5%	54.0%
	ComParE + PT features	65.4%	63.2%	65.5%	63.8%
ComParE 2014 baseline [22]		–	61.3%	–	61.5%

Examining the four datasets, we notice an important difference among them: in the Physical Load and in the Eating Condition corpus, we have to detect specific phenomena in the speech signal (i.e. heavy breathing and the sound of specific food types eaten), while in the Emotion and in the Cognitive Load corpora the task is to detect more subtle changes in the behavior

Please cite this article as: G. Gosztolya, Posterior-thresholding feature extraction for paralinguistic speech classification, Knowledge-Based Systems (2019) 104943, https://doi.org/10.1016/j.knosys.2019.104943.

Table 10
The combined results obtained by various approaches and authors on all the tasks of the Cognitive Load corpus.

Approach	Cross-validation		Test	
	Acc.	UAR	Acc.	UAR
ComParE functionals feature set (baseline)	68.0%	66.9%	65.9%	64.8%
Posterior-thresholding (PT) features	61.5%	59.5%	61.6%	59.5%
ComParE + PT features (MFCC)	68.1%	66.9%	66.7%	65.3%
ComParE + PT features (FBANK)	68.2%	67.0%	66.6%	65.2%
ComParE 2014 baseline [22]	–	63.2%	–	61.6%
VOQAL features + CART classifier [7]	–	66.5%	–	63.1%
High-level speech event analysis [72]	–	–	–	63.1%
ComParE + SCF + MFCC + SDC feature combination [73]	–	64.8%	–	63.7%
Prosodic and ASR-derived features [71]	–	77.5%	–	68.9%

of the speaker. Despite this, the emotion of the speaker could be detected at the frame level at an acceptable level, but the poor performance of the frame-level DNNs on the Cognitive Load corpus could be due to the task itself. Still, according to the experimental results, the Posterior-Thresholding Feature Extraction method could extract meaningful utterance-level features from the trends of the frame-level posterior values. On the Physical Load and Eating Condition datasets, however, the PTFE approach yielded accuracy scores close to the standard paralinguistic one, and the combination of the two strategies brought a significant improvement.

Tables 6 and 8 also contain the notable results achieved by other authors on the same dataset. (Recall that we had to repartition the Emotion corpus, so our results cannot be directly compared to those presented in [37,38]; another reason is that we focused on UAR, while these earlier studies only reported classification accuracy.) Examining the presented UAR scores, we may conclude that using the PTFE features in combination with the 6373-sized ComParE functionals feature set led to quite competitive results in all cases: the 74.7% achieved on the Physical Load dataset is only 0.7% lower than the Challenge–winner result of 75.4%, and the 79.7% UAR value obtained on the Eating Condition task also falls close to the 81.6% attained via Fisher vector analysis.

For the Cognitive Load dataset, unfortunately, comparing the results is not that straightforward, since this corpus contains three different tasks performed by the speakers, and the classification results are usually presented in a combined form instead of a per-task basis. However, two of these three tasks (the two variations of the Stroop test) contained so few utterances that we were unable to train frame-level DNN models (as described in Section 3.1) on them. Therefore for these two tasks we used the predictions obtained via the baseline approach, while for the *reading sentence* task we used the PTFE process proposed.

Table 10 lists the accuracy and UAR scores of the tested approaches and some notable scores present in the literature in this combined form. These scores show a similar tendency to those listed in Table 9: using the PT features led to competitive scores, and by combining this approach with the baseline one via late fusion, we were able to significantly exceed the performance of the baseline ComParE feature set. Both UAR scores obtained on the test set (65.3% and 65.2%, MFCC and FBANK frame-level feature sets, respectively) are significantly higher than most results published, which fell in the range 63.1%–63.7%.

Of course, just as in the case of the Physical Load and the Eating Condition datasets, the UAR score got via the PTFE approach is not the highest one published so far, as it lags behind the 68.9% score achieved by Van Segbroeck et al. [71], got by extracting prosodic and speech recognition-based features. In our opinion, however, the PTFE approach proposed in this study has two clear advantages over the procedures proposed in [9,26] and [71], despite providing slightly lower accuracy scores. Firstly, it can be easily realized by standard speech processing tools. The second and more important advantage of our approach is that, based on our experimental results, it is a task-independent procedure, since it led to quite good accuracy scores for four different computational paralinguistic datasets. (Previously, we had similar experiences with a less refined variation of the PTFE approach on two other corpora [28].) A further option might be to combine the PTFE predictions with those of other paralinguistic approaches. This, however, is clearly the subject of future work.

7. Conclusions

In the task of Automatic Speech Recognition (ASR), machine learning is usually done at the frame level using Deep Neural Networks. In computational paralinguistics, however, classification or regression takes place at the level of larger units like segments or whole utterances, and it relies on specific segment-level features. In this study we sought to fuse the two approaches: in the first step of the proposed Posterior-Thresholding Feature Extraction (PTFE) workflow, we train DNNs on standard frame-level features such as MFCCs. Then, in the second step, we extract utterance-level feature vectors from the frame-level DNN outputs (i.e. the posterior estimates), which, in the third step, are used to train an utterance-level classifier model. We tested our approach on four different computational paralinguistic datasets. The experimental results indicate that this method yields acceptable accuracy scores even on its own, but we managed to significantly exceed the baseline scores by combining our predictions with those got by using the standard paralinguistic approach. According to the results, the proposed PTFE workflow seems to be both language-independent and task-independent, as we got improvements on all four datasets, although the amount of improvement might depend on the type of the actual speech corpus. As regards ease-to-use, it can be easily realized using only standard speech recognition and machine learning tools, and it has no meta-parameter whose value needs to be set.

Acknowledgments

The Titan X graphics card used for this study was donated by the NVIDIA Corporation. This research was partially funded by the Ministry of Human Capacities, Hungary (grants 20391-3/2018/FEKUSTRAT and TUDFO/47138-1/2019-ITM), and by the National Research, Development and Innovation Office of Hungary via contract NKFIH FK-124413. G. Gosztolya was also funded by the János Bolyai Scholarship of the Hungarian Academy of Sciences and by the Hungarian Ministry of Innovation and Technology New National Excellence Program ÚNKP-19-4.

References

[1] B. Wheatley, G. Doddington, C. Hemphill, J. Godfrey, E. Holliman, J. McDaniel, D. Fisher, Robust automatic time alignment of orthographic transcriptions with unconstrained speech, in: Proceedings of ICASSP, 1992, pp. 533–536.
[2] J. Sato, S. Morishima, Emotion modeling in speech production using emotion space, in: Proceedings of ROMAN, 1996, pp. 472–477.
[3] R. Huber, E. Nöth, A. Batliner, J. Buckow, V. Warnke, H. Niemann, You BEEP machine – emotion in automatic speech understanding systems, in: Proceedings of TSD, 1998, pp. 223–228.
[4] F. Grèzes, J. Richards, A. Rosenberg, Let me finish: Automatic conflict detection using speaker overlap, in: Proceedings of Interspeech, 2013, pp. 200–204.
[5] M.-J. Caraty, C. Montacié, Detecting speech interruptions for automatic conflict detection, in: Conflict and Multimodal Communication, Springer International Publishing, 2015, pp. 377–401, Ch. 18.
[6] M. van Segbroeck, R. Travadi, C. Vaz, J. Kim, M.P. Black, A. Potamianos, S.S. Narayanan, Classification of cognitive load from speech using an i-vector framework, in: Proceedings of Interspeech, Singapore, 2014, pp. 671–675.
[7] M. Huckvale, Prediction of cognitive load from speech with the VOQAL voice quality toolbox for the InterSpeech 2014 Computational Paralinguistics Challenge, in: Proceedings of Interspeech, 2014, pp. 741–745.
[8] G. Gosztolya, T. Grósz, R. Busa-Fekete, L. Tóth, Detecting the intensity of cognitive and physical load using AdaBoost and Deep Rectifier Neural Networks, in: Proceedings of Interspeech, Singapore, 2014, pp. 452–456.
[9] H. Kaya, T. Özkaptan, A.A. Salah, F. Gürgen, Canonical correlation analysis and Local Fisher Discriminant Analysis based multi-view acoustic feature reduction for physical load prediction, in: Proceedings of Interspeech, Singapore, 2014, pp. 442–446.
[10] D. Bone, M.P. Black, M. Li, A. Metallinou, S. Lee, S.S. Narayanan, Intoxicated speech detection by fusion of speaker normalized hierarchical features and GMM supervectors, in: Proceedings of Interspeech, 2011, pp. 3217–3220.
[11] C. Montacié, M.-J. Caraty, Combining multiple phoneme-based classifiers with audio feature-based classifier for the detection of alcohol intoxication, in: Proceedings of Interspeech, 2011, pp. 3205–3208.
[12] I. Hoffmann, D. Németh, C. Dye, M. Pákáski, T. Irinyi, J. Kálmán, Temporal parameters of spontaneous speech in Alzheimer's disease, Int. J. Speech-Lang. Pathol. 12 (1) (2010) 29–34.
[13] J.-R. Orozco-Arroyave, J. Arias-Londono, J. Vargas-Bonilla, E. Nöth, Analysis of speech from people with Parkinson's disease through nonlinear dynamics, in: Proceedings of NoLISP, 2013, pp. 112–119.
[14] G. Kiss, M.G. Tulics, D. Sztahó, K. Vicsi, Language independent detection possibilities of depression by speech, in: Proceedings of NoLISP, 2016, pp. 103–114.
[15] J. Weiner, C. Herff, T. Schultz, Speech-based detection of Alzheimers Disease in conversational German, in: Proceedings of Interspeech, San Francisco, CA, USA, 2016, pp. 1938–1942.
[16] B. Schuller, S. Steidl, A. Batliner, The Interspeech 2009 emotion challenge, in: Proceedings of Interspeech, Brighton, UK, 2009, pp. 312–315.
[17] L. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.
[18] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, R.C. Williamson, Estimating the support of a high-dimensional distribution, Neural Comput. 13 (7) (2001) 1443–1471.
[19] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Process. Mag. 29 (6) (2012) 82–97.
[20] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente, S. Kim, The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism, in: Proceedings of Interspeech, Lyon, France, 2013, pp. 148–152.
[21] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, Y. Zhang, The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load, in: Proceedings of Interspeech, 2014, pp. 427–431.
[22] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J.R. Orozco-Arroyave, E. Nöth, Y. Zhang, F. Weninger, The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition, in: Proceedings of Interspeech, 2015, pp. 478–482.
[23] G. Gosztolya, A. Beke, T. Neuberger, L. Tóth, Laughter classification using Deep Rectifier Neural Networks with a minimal feature subset, Arch. Acoust. 41 (4) (2016) 669–682.
[24] R. Brueckner, M. Schmitt, M. Pantic, B. Schuller, Spotting social signals in conversational speech over IP: A deep learning perspective, in: Proceedings of Interspeech, 2017, pp. 2371–2375.
[25] H. Inaguma, K. Inoue, M. Mimura, T. Kawahara, Social signal detection in spontaneous dialogue using bidirectional LSTM-CTC, in: Proceedings of Interspeech, 2017, pp. 1691–1695.
[26] H. Kaya, A.A. Karpov, A.A. Salah, Fisher Vectors with cascaded normalization for paralinguistic analysis, in: Proceedings of Interspeech, 2015, pp. 909–913.
[27] G. Gosztolya, L. Tóth, DNN-based feature extraction for conflict intensity estimation from speech, IEEE Signal Process. Lett. 24 (12) (2017) 1837–1841.
[28] G. Gosztolya, R. Busa-Fekete, T. Grósz, L. Tóth, DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification, in: Proceedings of Interspeech, Stockholm, Sweden, 2017, pp. 3522–3526.
[29] S. Pancoast, M. Akbacak, Bag-of-Audio-Words approach for multimedia event classification, in: Proceedings of Interspeech, Portland, OR, USA, 2012, pp. 2105–2108.
[30] F.B. Pokorny, F. Graf, F. Pernkopf, B.W. Schuller, Detection of negative emotions in speech signals using bags-of-audio-words, in: Proceedings of ACII, 2015, pp. 1–5.
[31] M. Schmitt, F. Ringeval, B. Schuller, At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech, in: Proceedings of Interspeech, San Francisco, CA, USA, 2016, pp. 495–499.
[32] A. Ouahabi (Ed.), Signal and Image Multiresolution Analysis, Polytech'Tours, Paris, France, 2012.
[33] C.P. Chan, P.C. Ching, T. Lee, Noisy speech recognition using de-noised multiresolution analysis acoustic features, J. Acoust. Soc. Am. 110 (2001) 2567–2574.
[34] K. Wang, Time-frequency feature representation using multi-resolution texture analysis and acoustic activity detector for real-life speech emotion recognition, Sensors 15 (2015) 1458–1478.
[35] G.y. Kovács, L. Tóth, D. van Compernolle, S. Ganapathy, Increasing the robustness of CNN acoustic models using ARMA spectrogram features and channel dropout, Pattern Recognit. Lett. 100 (2017) 44–50.
[36] B. Schuller, F. Friedmann, F. Eyben, The Munich Biovoice Corpus: Effects of physical exercising, heart rate, and skin conductance on human speech production, in: Proceedings of LREC, 2014, pp. 1506–1510.
[37] D. Sztahó, V. Imre, K. Vicsi, Automatic classification of emotions in spontaneous speech, in: Proceedings of COST 2102, Budapest, Hungary, 2011, pp. 229–239.
[38] K. Vicsi, D. Sztahó, Recognition of emotions on the basis of different levels of speech segments, J. Adv. Comput. Intell. Intell. Inform. 16 (2) (2012) 335–340.
[39] S. Hantke, F. Weninger, R. Kurle, F. Ringeval, A. Batliner, A.E.-D. Mousa, B. Schuller, I hear you eat and speak: Automatic recognition of eating condition and food type, use-cases, and impact on ASR performance, PLoS One (2016) 1–24.
[40] T.F. Yap, Speech Production under Cognitive Load: Effects and Classification (Ph.D. thesis), University of New South Wales, 2012.
[41] J.R. Stroop, Studies of interference in serial verbal reactions, J. Exp. Psychol. 18 (6) (1935) 643–662.
[42] S. Young, G. Evermann, M.J.F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book, Cambridge University Engineering Department, Cambridge, UK, 2006.
[43] L. Xu, A. Krzyzak, C. Suen, Methods for combining multiple classifiers and their applications to handwriting recognition, IEEE Trans. Syst. Man Cybern. 22 (3) (1992) 418–435.
[44] D. Yu, L. Deng, Automatic Speech Recognition: A Deep Learning Approach, Springer Publishing Company, 2014.
[45] R.A. Schowengerdt, Remote Sensing: Models and Methods for Image Processing, Academic Press, Orlando, FL, USA, 2006.
[46] S. Molau, M. Pitz, H. Ney, Histogram based normalization in the acoustic feature space, in: Proceedings of ASRU, Madonna di Campiglio, Italy, 2001, pp. 1–4.
[47] P. Hiremath, S. Shivashankar, Wavelet based co-occurrence histogram features for texture classification with an application to script identification in a document image, Pattern Recognit. Lett. 29 (2008) 1182–1189.
[48] L. Heutte, T. Paquet, J. Moreau, Y. Lecourtier, C. Olivier, A structural/statistical feature based vector for handwritten character recognition, Pattern Recognit. Lett. 19 (1998) 629–641.
[49] F. Alegria, A. da Cruz Serra, Influence of frequency errors in the variance of the cumulative histogram [in adc testing], IEEE Trans. Instrum. Meas. 50 (2001) 461–464.
[50] B. Milde, C. Biemann, Using representation learning and out-of-domain data for a paralinguistic speech task, in: Proceedings of Interspeech, Dresden, Germany, 2015, pp. 904–908.
[51] H.-S. Lee, Y. Tsao, C.-C. Lee, H.-M. Wang, W.-C. Lin, W.-C. Chen, S.-W. Hsiao, S.-K. Jeng, Minimization of regression and ranking losses with shallow neural networks on automatic sincerity evaluation, in: Proceedings of Interspeech, San Francisco, CA, USA, 2016, pp. 2031–2035.

Please cite this article as: G. Gosztolya, Posterior-thresholding feature extraction for paralinguistic speech classification, Knowledge-Based Systems (2019) 104943, https://doi.org/10.1016/j.knosys.2019.104943.

[52] G. Gosztolya, T. Grósz, R. Busa-Fekete, L. Tóth, Determining native language and deception using phonetic features and classifier combination, in: Proceedings of Interspeech, San Francisco, CA, USA, 2016, pp. 2418–2422.
[53] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J.K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini, The Interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language, in: Proceedings of Interspeech, San Francisco, CA, USA, 2016, pp. 2001–2005.
[54] B. Schuller, S. Steidl, A. Batliner, S. Hantke, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A.S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, S. Zafeiriou, The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, Cold & Snoring, in: Proceedings of Interspeech, 2017, pp. 1–5.
[55] G. Gosztolya, T. Grósz, G.y. Szaszák, L. Tóth, Estimating the sincerity of apologies in speech by DNN rank learning and prosodic analysis, in: Proceedings of Interspeech, San Francisco, CA, USA, 2016, pp. 2026–2030.
[56] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proc. AISTATS, 2010, pp. 249–256.
[57] L. Tóth, Phone recognition with hierarchical Convolutional Deep Maxout Networks, EURASIP J. Audio Speech Music Process. 2015 (25) (2015) 1–13.
[58] T. Grósz, R. Busa-Fekete, G. Gosztolya, L. Tóth, Assessing the degree of nativeness and Parkinson's condition using Gaussian processes and Deep Rectifier Neural Networks, in: Proceedings of Interspeech, Dresden, Germany, 2015, pp. 1339–1343.
[59] C. Zhang, P.C. Woodland, A general Artificial Neural Network extension for HTK, in: Proceedings of Interspeech, Dresden, Germany, 2015, pp. 3581–3585.
[60] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (2011) 1–27.
[61] G. Gosztolya, Conflict intensity estimation from speech using greedy forward-backward feature selection, in: Proceedings of Interspeech, Dresden, Germany, 2015, pp. 1339–1343.
[62] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of German emotional speech, in: Proceedings of Interspeech, Lisbon, Portugal, 2005, pp. 1517–1520.
[63] S. Steidl, Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech, Logos Verlag, Berlin, 2009.
[64] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: The Munich versatile and fast open-source audio feature extractor, in: Proceedings of ACM Multimedia, 2010, pp. 1459–1462.
[65] M. Schmitt, B. Schuller, openXBOW – Introducing the Passau open-source crossmodal Bag-of-Words toolkit, J. Mach. Learn. Res. 18 (2017) 1–5.
[66] G. Gosztolya, L. Szilágyi, Application of fuzzy and possibilistic *c*-means clustering models in blind speaker clustering, Acta Polytech. Hungar. 12 (7) (2015) 41–56.
[67] G. Gosztolya, L. Tóth, A feature selection-based speaker clustering method for paralinguistic tasks, Pattern Anal. Appl. 21 (1) (2018) 193–204.
[68] K.J. Han, S. Kim, S.S. Narayanan, Strategies to improve the robustness of Agglomerative Hierarchical Clustering under data source variation for speaker diarization, IEEE Trans. Audio Speech Lang. Process. 16 (8) (2008) 1590–1601.
[69] W. Wang, P. Lu, Y. Yan, An improved hierarchical speaker clustering, Acta Acust. 33 (1) (2008) 9–14.
[70] A. Nguyen, J. Yosinski, J. Clune, Deep Neural Network are easily fooled: High confidence predictions for unrecognizable images, in: Proceedings of CVPR, Boston, MA, USA, 2015, pp. 427–436.
[71] M. van Segbroeck, R. Travadi, C. Vaz, J. Kim, M.P. Black, A. Potamianos, S.S. Narayanan, Classification of cognitive load from speech using an i-vector framework, in: Proceedings of Interspeech, Singapore, 2014, pp. 671–675.
[72] C. Montacié, M.-J. Caraty, High-level speech event analysis for cognitive load classification, in: Proceedings of Interspeech, 2014, pp. 731–735.
[73] J.M.K. Kua, V. Sethu, P. Le, E. Ambikairajah, The UNSW submission to INTERSPEECH 2014 ComParE Cognitive Load challenge, in: Proceedings of Interspeech, 2014, pp. 746–750.

Please cite this article as: G. Gosztolya, Posterior-thresholding feature extraction for paralinguistic speech classification, Knowledge-Based Systems (2019) 104943, https://doi.org/10.1016/j.knosys.2019.104943.