# KOLMOGOROV-SMIRNOV TEST FOR FEATURE SELECTION IN EMOTION RECOGNITION FROM SPEECH

*Alexei Ivanov and Giuseppe Riccardi*

Department of Information Engineering and Computer Science
University of Trento, Italy
alexei_v_ivanov@ieee.org, riccardi@disi.unitn.it

## ABSTRACT

Automatic emotion recognition from speech is limited by the ability to discover the relevant predicting features. The common approach is to extract a very large set of features over a generally long analysis time window. In this paper we investigate the applicability of two-sample Kolmogorov-Smirnov statistical test (KST) to the problem of segmental speech emotion recognition. We train emotion classifiers for each speech segment within an utterance. The segment labels are then combined to predict the *dominant* emotion label. Our findings show that KST can be successfully used to extract statistically relevant features. KST criterion is used to optimize the parameters of the statistical segmental analysis, namely the window segment size and shift. We carry out seven binary class emotion classification experiments on the Emo-DB and evaluate the impact of the segmental analysis and emotion-specific feature selection.

***Index Terms—*** feature selection, emotion recognition, Kolmogorov-Smirnov statistics

## 1. INTRODUCTION

Current automated systems for emotion recognition from speech have shown a significant progress for instances of *clearly manifested acted prototypical basic emotions* (e.g. anger, joy) [1, 2]. Alternatively to categorical labeling, the emotional state may be characterized in terms of several real-valued parameters (typically arousal and valence) [1]. As these systems are based on statistically-estimated features, a sufficient evidence ($\sim$ 2 seconds of speech) has to be collected for reliable operation. However, these systems perform much worse in the recognition of real-life spontaneous emotional manifestations.

In the case of real-life spontaneous spoken interaction, the instantaneous affective state of interlocutors "is a mixture of emotion, attitude, mood, interpersonal stance, often in response to multi-trigger events (both internal and external) occurring at different times" [3]. While being more localized in time compared to the whole utterance, these cues are being injected into speech each at a certain time instance to produce a *dominant* affected utterance. The conversational context is important for the process of meaning attribution (grounding) of the observed interactive cues [4] and the attribution of a particular hypothesized affective state. For affect recognition technology to become more robust in real tasks a more sophisticated detection system has to be employed. One which is capable of detection individual affective cues within smaller analysis intervals and proper contextualized cue interpretation.

The approach of emotion recognition from sub-intervals was tried in [5] and it has proven to be advantageous in combination with the traditional large-analysis span feature vector computation. However, a systematic way to select the optimal sub-interval split was not given. Besides, the sub-intervals were strictly non-overlapping. Detection of the emotion-specific phoneme-conditioned cues was explored in [6] and proved to be informative for predicting the emotional label for the whole utterance. However, only low-level spectral features (MFCCs) were evaluated. A recent paper [7] has explored a larger sliding analysis window of 1 sec. The standard feature vector of Interspeech'2009 Emotional Speech Challenge [8] was universally used. An improvement in comparison to a baseline (computing a unique feature vector from the whole utterance) was recorded. The improvement was varying with a particular choice of heuristics for the global emotional label generation. A problem of feature selection was addressed in [9]. However the approach adopted there was an exhaustive search for the best feature combination while retraining SVM classifier on the whole database.

In this paper we explore the predictive power of the two-sample Kolmogorov-Smirnov statistical test on the individual features towards selection of the optimal width and shift of the statistical analysis interval, as well as the set of features for given optimal width and shift. Recently the two-sample KST has been found useful in comparative feature evaluation for emotion recognition [10]. However there are no reports on KST–based feature selection.

## 2. KOLMOGOROV-SMIRNOV STATISTICAL TEST FOR FEATURE SELECTION

It is possible to use a two-sample KST to assess the similarity of empirically defined distributions of random variables. KST aims at rejecting (at the specified level of significance

$p$) the null-hypothesis $H_0$ that two random variables have identical distributions. Essentially for a given pair of random variables $X$ and $Y$ the Kolmogorov-Smirnov statistics $D_{X,Y}$ is the largest observed discrepancy between the estimated cumulative distributions through-out the sample space (i.e. $\forall z \in (-\infty, \infty)$):

$$D_{X,Y} = \sup_{\forall z \in (-\infty,\infty)} |\hat{P}(X \leq z) - \hat{P}(Y \leq z)|. \quad (1)$$

Here "sup" refers to a supremum operation, i.e. a choice of the largest operand value. The associated significance $p$ is obtained as a probability to see an observed value of $D_{X,Y}$ under $H_0$ while drawing the random samples of $X$ and $Y$.

The usefulness of KST in application to feature selection comes from the absence of the explicit analytical assumption on the form of the distributions of $X$ and $Y$. Given a classical binary classification task one might search for features which individually violate $H_0$ at a significance level $p$ when samples are coming from the binary classes ($C = 0$ and $C = 1$).

In the uni-variate case, when

$$P(X \leq z) = 1 - P(X > z) \quad (2)$$

and, thus, according to (1), $D_{X,Y}$ is invariant in respect to orientation of the sample space, i.e. whether we traverse it from $-\infty$ to $\infty$ or the other way around. Substitution of the '$\leq$' sign with a '$<$' sign during order inversion is not essential for the definition of the Kolmogorov-Smirnov statistics.

Unfortunately there is no straightforward generalization of KST to multivariate analysis. E.g. for the bi-variate case the following statement holds true

$$\begin{aligned} P(X_1 \leq z_1, X_2 \leq z_2) = 1 - P\left(X_1 > z_1, X_2 > z_2\right) - \\ -P(X_1 > z_1, X_2 \leq z_2) - P(X_1 \leq z_1, X_2 > z_2)\,, \end{aligned} \quad (3)$$

which implies that according to (1) the estimates of $D_{X,Y}$ are no longer required to be equal regardless of the direction of the sample space traversal.

For a general $M$-variate case there are $2^M - 1$ independent ways to re-arrange the sample space and, as a result, obtain possibly different values for $D_{X,Y}$ and even have different outcomes of the test. See [11] for a detailed discussion on existing multi-variate generalizations. Besides, reliable estimation of the multi-variate distribution requires a progressively larger amount of empirical evidence with the growth of the sample space dimensionality.

## 3. CORPORA DESCRIPTION

All experiments described in the present paper were performed with the database of acted German emotional speech (Emo-DB) [12]. The database contains utterances from 10 native speakers of German. In each utterance the speakers enact one of ten prototypical emotion state: anger, boredom, disgust, fear, happiness, neutral, sadness. A subset of the collected speech containing 494 utterances, which passes the inter-annotator agreement threshold, is used for experiments.

## 4. EXPERIMENTS

The baseline feature set used in our experiments, described in this paper, corresponds to the one suggested in Interspeech 2011 Speaker State Challenge [13]. In total there are 4368 features. The whole set consists of a detailed statistical description of the basic speech features. According to the approach adopted in openSMILE [14] the feature extraction is a two-tier process. First, the basic features are being extracted within a uniform observation window of 10 msec, then statistics are drawn from a larger statistical analysis window. Variation of the analysis window width and shift refer to the latter (statistical analysis) window.

### 4.1. Kolmogorov-Smirnov test predictions

KST is applied to each individual feature in the collection with the aim to reject the hypothesis that it is statistically relevant for a target emotion label. For each emotion label we compute the binary data split where the label is (not) annotated. The hypothesis to reject is that distributions in both parts of the binary data split are identical (e.g. in samples pertaining to 'a class' as well as the remaining part of the data, which is tagged as 'not-a-class'). The property of the two-sample KST is that both compared distributions are defined empirically and are not required to be represented by the same number of samples.

There are two major ways to present a joint outcome of individual feature tests:1) fix the parameter value (e.g. analysis interval, shift, etc.) and see how many features "survive" a test at a given significance level; 2) plot a histogram across the parameter value space and put features into the bins, corresponding to the most advantageous parameter value for them.

An example of the first type of analysis is presented in Fig. 1. In each of the plots there is a family of curves, that correspond to a different choice of the statistical analysis window shift and the used significance level. A property of these curves is that with a proper choice of the combination of shift and significance level, they tend to overlap, while not being exactly the same. It is not only the size of the sets being similar, the overlap of the sets typically reaches a high level of $\sim 98\%$. From the practical standpoint this property may find an application in predicting the outcome of the computationally expensive KST (when shift value is small) by the KST results with the shift being larger.

An alternative type of analysis would be to plot a histogram and answer the question of how many features are the most relevant at a particular analysis interval. An example of such plot is given on Fig. 2. Here we take a $p$-value, associated with the rejection of $H_0$ in KST trial for each individual feature, as a measure of that feature reliability. All features, that do not pass a reliability test at the level of $p = 0.05$ are deemed as completely unreliable for a given task. Those are binned in a separate category. As one can see from the plot, there is a large share of features, which have the maximum of their reliability at the shortest analysis interval. From the inspection of the selected features we find that simple statistics

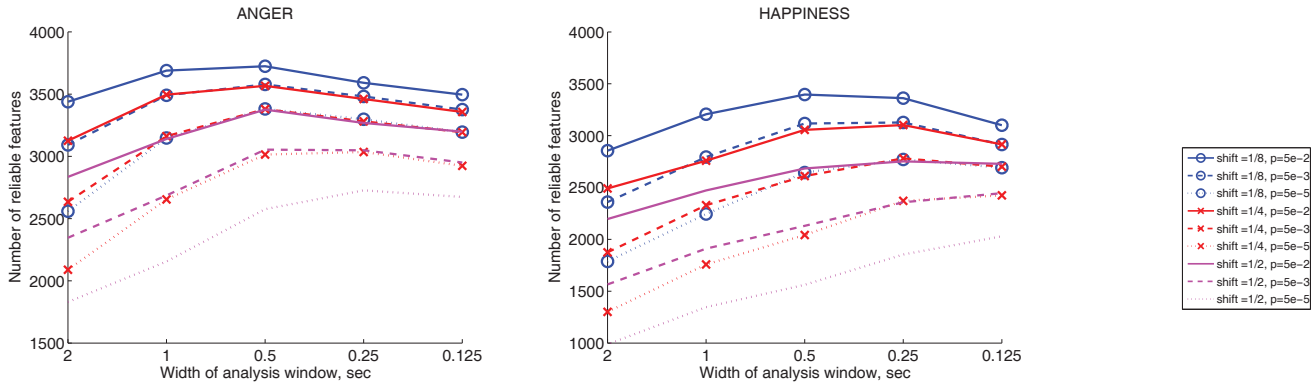


**Fig. 1**. KST prediction for number of relevant features for a given fixed analysis interval length. Curves are shown for the 'ANGER' and 'HAPPINESS' labels.

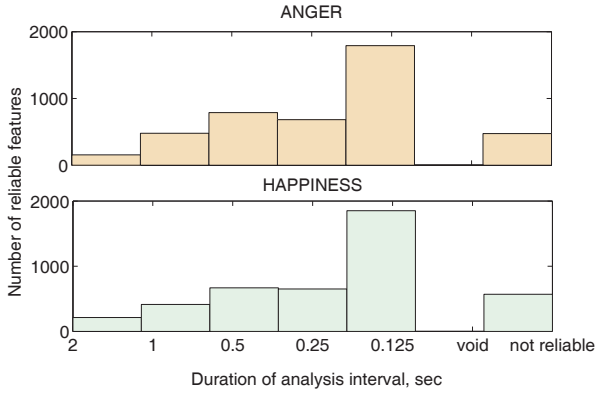


**Fig. 2**. Histogram of the distribution of the optimal analysis window size for features from Interspeech'2011 collection.

such as mean, variance, etc. tend to group in that plot interval.

In general there is a large contrast between the picture of Fig. 2 and the number of relevant features for a given fixed analysis interval length (Fig. 1). This observed difference suggests that a multi-rate feature acquisition, local classification and fusion of local decisions into a global one should be advantageous strategy for emotion recognition from speech.

### 4.2. Recognition Experiments

The speech data is split 10-ways for cross-validation by leaving one speaker out (LOSO). Thus each of the 10 test sets contains the data coming from a single speaker that was not present in the corresponding training set. For statistical modeling the state-of-the-art emotion recognition system has been used. The system consists of an openSMILE-based feature extraction [14] and the boostexter classifier [15]. In separate experiments on the same database but with a larger feature set this system has been attaining the best performance of $WA \sim 85\%$ (LOSO).

The reported baseline system (Table 1 experiment 1) strictly follows a conventional method of classification of the unique feature vector, that is computed by openSMILE from the whole utterance of variable duration.

The whole system performance is evaluated with the help of two widely accepted figures of merit, the first is weighted accuracy (WA) and the second is unweighted accuracy (UA):

$$WA = \frac{\sum_{\forall k} N_{Corr_k}}{\sum_{\forall k} N_{Tot_k}}; \quad UA = \frac{1}{K} \sum_{\forall k} \frac{N_{Corr_k}}{N_{Tot_k}}. \qquad (4)$$

Here it is assumed that $k = \overline{1..K}$ enumerates the labels; $N_{Corr_k}$ stands for number of correctly recognized instances of a given label; $N_{Tot_k}$ is a total number of instances of a that label.

#### *4.2.1. The Optimal Analysis Interval Length and Shift*

In majority of emotion labels, the largest number of relevant features is observed while the analysis window size was approximately from 1 to 0.5 sec. In the recognition experiment however the maximum performance is observed for the analysis window size being 1 sec. See Table 1 for details. The KST prediction can be used as an approximate indication for selection of the advantageous analysis interval length.

**Table 1**. Recognition experiment for different analysis window length.'Exp.'- experiment ID; 'Wind.Size' - size of the statistical analysis window; 'Shift' - shift in time between the adjacent windows; 'WA' - weighted accuracy; 'UA' - unweighted accuracy.

| Exp. | Wind.Size | Shift | WA | UA |
|---|---|---|---|---|
| 1 | whole utt. | N/A | 76.11% | 71.82% |
| 2 | 1.5 sec | 1/8 | 74.79% | 71/69% |
| **3** | **1.0 sec** | **1/8** | **76.32%** | **73.60%** |
| 4 | 0.5 sec | 1/8 | 70.04% | 65.45% |
| 5 | 1.0 sec | 1/4 | 71.05% | 69.25% |

KST analysis suggests, that a greater overlap of the analysis windows results in a greater amount of relevant features

for classification. The recognition experiments are in agreement with this prediction. To illustrate this fact the experiment 5 form Table 1 summarizes results of the recognition experiment, when the analysis frame shift is increased two times compared to the experiment 3.

#### *4.2.2. Feature Pruning*

Two recognition experiments have been performed to explore the ability of KST to drive feature selection in the emotion recognition task. The first experiment aimed at measuring

**Table 2**. Recognition experiment with reduced feature sets.'Em'- emotion label in binary classification task; 'A'- anger; 'B'- boredom; 'D'- disgust; 'F'- fear; 'J'- joy; 'N'- neutral; 'S'- sadness; 'Cr'- emotion label cardinality within the test set; 'BL' - baseline feature set; 'Red' - reduced feature set; 'WA' - weighted accuracy; 'UA' - unweighted accuracy; '% feat.' - percent of features retained in the reduced set.

| Em (Cr) | BL WA | BL UA | Red WA | Red UA | % feat. |
|---|---|---|---|---|---|
| Feature vector is computed from the whole utterance | | | | | |
| A(127) | 90.08% | 88.18% | **90.49%** | **88.70%** | 68.06% |
| B(79) | 96.36% | 90.14% | **96.96%** | **92.04%** | 56.98% |
| D(38) | 95.75% | 78.48% | **96.56%** | **81.25%** | 35.07% |
| F(55) | 92.91% | 72.95% | **94.74%** | **80.34%** | 41.90% |
| J(64) | 88.26% | 62.00% | 87.65% | 60.99% | 50.66% |
| N(78) | 92.71% | 84.21% | **93.32%** | **85.10%** | 49.61% |
| S(53) | 97.17% | 90.11% | 96.76% | 89.89% | 71.22% |
| Feature vector is computed from the 1.0 sec window, shift 1/8 | | | | | |
| A(127) | 91.09% | 88.34% | **91.09%** | **88.60%** | 80.03% |
| B(79) | 93.52% | 82.82% | **93.72%** | **83.45%** | 73.15% |
| D(38) | 94.13% | 70.29% | **94.74%** | 69.41% | 53.02% |
| F(55) | 91.09% | 62.39% | **91.50%** | **63.41%** | 49.40% |
| J(64) | 88.87% | 61.69% | **89.07%** | **63.13%** | 65.06% |
| N(78) | 90.28% | 74.44% | 89.47% | 69.79% | 65.52% |
| S(53) | 94.74% | 84.60% | 94.33% | 83.55% | 82.67% |

the impact of pruning unreliable (according to KST) features from the utterance-level description. The second is the same measurement for the segmental recognizer with 1 sec observation window and 125 msec shift. There are 7 binary classifiers (one for each emotional label), each having its own set of features, specific for a given emotion label. The features which are deemed as not-reliable at a significance value (p=$5e^{-2}$) are excluded. The feature pruning is quite severe, sometimes the cut exceeds 50% of the feature set.

System's performance in both experiments is reported in Table 2. KST feature pruning has a positive effect onto the system performance at the level of individual binary classifiers, in majority of the cases the performance has increased while the feature set has been reduced. Amount of features surviving the pruning process may be regulated by a different choice of the significance level $p$. Larger $p$ values allow to retain more features.

## 5. CONCLUSIONS

In this paper we have applied the two-sample Kolmogorov-Smirnov test to feature analysis of emotion recognition from speech and explored its predictions. KST has been able to correctly predict the effect of the decreased analysis window shift onto recognition performance. It is also very successful in predicting which features are safe to prune from the vector. We also have found that according to KST a multi-rate feature extraction shall be advantageous for emotion recognition.

## 6. ACKNOWLEDGMENT

This work was partially supported by the Livememories project funded by Autonomous Province of Trento.

## 7. REFERENCES

[1] B. Schuller, B. Vlasenko, F. Eybena, G.Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. of ASRU'2009*, Merano, Italy, Dec. 2009.

[2] M. El Ayadi, M.S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classication schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[3] L. Devillers and eds. N. Campbell, "Special issue of computer speech and language on affective speech in real-life interactions," *Computer Speech and Language*, vol. 25, no. 1, 2011.

[4] G. Riccardi and D. Hakkani-Tür, "Grounding emotions in human-machine conversational systems," *Lect. notes in Comp. Sc.*, pp. 144–154, 2005.

[5] B. Schuller and G. Rigoll, "Timing levels in segment-based speech emotion recognition," in *Proc. of Interspeech'2006-ICSLP*, Pittsburgh, USA, Sept. 2006.

[6] C. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *Proc. of ICSLP'2004*, Jeju Island, Korea, Oct. 2004.

[7] J. H. Jeon, R. Xia, and Y. Liu, "Sentence level emotion recognition based on decisions from subsentnce segments," in *Proc. of ICASSP'2011*, Prague, Czech Rep., May 2011.

[8] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. Interspeech'2009*, Brighton, UK, Aug. 2009.

[9] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit - searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech & Language*, vol. 25, no. 1, pp. 4–28, 2011.

[10] H. Patro, G. Senthil Raja, and S. Dandapat, "Statistical feature evaluation for classification of stressed speech," *Int. J. of Sp. Tech.*, vol. 10, pp. 143–152, 2007.

[11] R.H.C. Lopes, I. Reid, and P.R. Hobson, "The two-dimensional Kolmogorov-Smirnov test," in *Proc. XI Int. Workshop on Adv. Computing and Analysis Tech. in Physics Res.*, April 2007.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. of Interspeech'2005*.

[13] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The interspeech 2011 speaker state challenge," in *Proc. Interspeech'2011*, Florence, Italy, Aug. 2011.

[14] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (MM), Florence, Italy*, 2010, pp. 1459–1462.

[15] R. Schapire and Y. Singer, "Boostexter: A boosting- based system for text categorization," 2000, pp. 135–168.