A Frame Pruning Approach for Paralinguistic Recognition Tasks

Johannes Wagner[1], *Florian Lingenfelser*[1], *Elisabeth André*[1]

[1]University of Augsburg, Lab for Human Centered Multimedia, Germany

[wagner,lingenfelser,andre]@hcm-lab.de

Abstract

In conventional paralinguistic classification approaches, information gained by low level features is described over broad segments (like whole turns) via statistical functionals. This procedure presumes meaningful information to be embodied within the whole segment. This assumption may be misleading if distinctive cues within a sample are surrounded by non-meaningful information or noise. In this case it would surely be beneficial to keep only parts of the sample that are most relevant for the recognition task. In this paper we propose a novel cluster-based approach, which aims at identifying frames likely to carry distinctive information. Evaluation is done within the INTERSPEECH 2012 Speaker Trait Challenge. Results show that under certain configurations frame pruning in fact leads to an improvement in recognition accuracy. On the observed corpus most stable improvements were achieved at a frame drop of 4-8%.

Index Terms: paralinguistic recognition, frame pruning, personality traits

1. Introduction

Paralanguage is known as an important carrier of non-verbal information. It gives off cues about our feelings and attitudes, while it also includes signs related to personality traits. Automatic recognition of such behaviour, though an important step towards a more intuitive human-computer interaction, turns out to be challenging, especially when applied under realistic settings.

The common approach to tackle a paralinguistic recognition problem is by applying statistical functionals to low level descriptors gained over broad time segments. Especially when dealing with acted speech, this technique has proven to be well suited as actors are used to exaggerated accentuation over the whole turn. In natural speech, however, it becomes less likely that meaningful information is embodied throughout an observed segment. In many cases, broad segments of the signal simply do not carry any meaningful information – real paralinguistic cues are included in few emphasised parts. Likewise, outside the lab there is an increased chance that an utterance will be partly overlaid by noise. At this point recognition of paralinguistic traits could be enhanced, when we manage to isolate significant parts of speech and ignore non-meaningful periods.

To this end we propose a novel cluster-based approach, which aims at identifying frames likely to carry distinctive information and use only those during classification. It is inspired by a technique applied in speaker verification tasks. To find the most likely speaker in speaker identification tasks, it is common practise to average frame scores over the whole test utterance. However, Besacier et al. [1] decided to apply a hard threshold to sum only frame scores that are most likely to identify a speaker, which led to a significant improvement of identification rate. Transferred to our problem, we want to identify frames that carry meaningful paralinguistic information for our recognition task. As test bed we take the INTERSPEECH 2012 Speaker Trait Challenge and focus on the *Personality Sub-Challenge* [2]. We will start with an introduction to the applied approach. Afterwards we present our results and try to answer the question if frame pruning can improve recognition accuracies and how the amount of pruned frames connect to it.

2. Methodology

In the following we will describe the suggested pruning approach. First, training samples are divided in short segments and converted into a compact representation. Afterwards, meaningful segments are separated from non-meaningful ones. Only meaningful segments are used during classification (see Figure 1 for a schematic). In this ways, we hope to avoid flattening effects, which may occur when parts of the signals with distinctive information are overlaid by meaningless or noisy parts in its surrounding.

2.1. Segmentation

The first step of the algorithm concerns the segmentation of the input samples into smaller parts. It can be based on a fixed length, *i. e.* a certain frame size, or exploit some intrinsic property of the signal, *e. g.* separation into voiced and unvoiced parts.

We decided in favour of a frame-based segmentation, in order to use low-level features (LLF) suggested by the

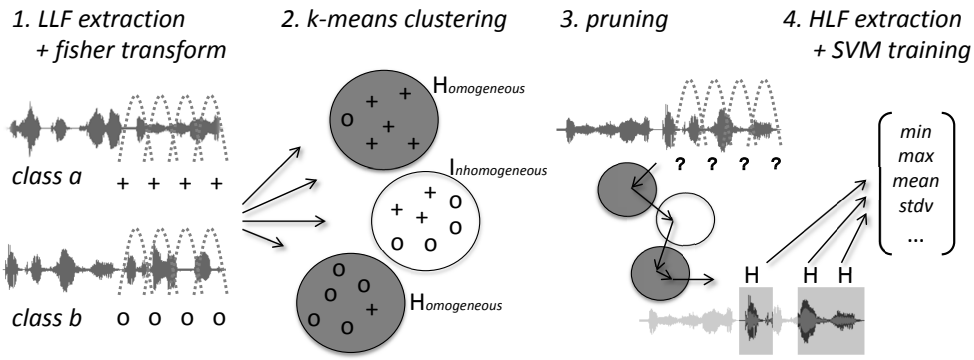

Figure 1: First, low-level feature are extracted for each frame and Fisher projection is applied (1). Next, transformed frames are clustered using K-Means. If the majority of frames inside a cluster belong to the same class we regard the cluster as *homogeneous*, otherwise we mark the cluster as *inhomogeneous* (2). Low-level feature vectors are then pruned by cutting out frames belonging to *inhomogeneous* clusters (3). Finally, high-level features are extracted from the pruned samples and used to train a SVM classifier (4).

organisers of the challenge. Therefore, we re-calculate features at a frame step of 0.01s using TUMs open-source openSMILE feature extractor [3], but remove extraction of statistical features. The list of the low-level-descriptors can be found in Table 4 in [2]. Together with delta regression coefficients, for each frame 129 low-level features are extracted and linked with the class label of the according sample.

Afterwards we apply a Fisher projection [4] on the frames in the training set in order to find a transformation into a feature space with less dimensions (in our case 2, as we have a 2-class problem). Though, feature reduction is not an essential part for our algorithm it significantly speeds up the following clustering step. This is for two reasons: Firstly, a reduced dimensionality leads to a speed up by itself. And secondly, since Fisher projection uses *Linear Discriminant Analysis* (LDA) to minimize the variance within classes and maximize the variance between classes, a better class separation is obtained for the target space, which increases the probability to find *homogeneous* regions at a smaller number of clusters.

2.2. Clustering

Now, we want to locate frames that carry distinctive class information. With respect to the transformed feature space, this concerns frames that are surrounded by frames of the same class, whilst far from frames of other classes. To this end, we apply K-Means clustering in order to group frames into well separated clusters[1] – yet independent of any class affiliation. Frames in the same cluster are alike, those belonging to different clusters show differences.

[1]K-Means clustering is a popular method because of its simplicity and stable performance – n observations are partitioned into k clusters in which each observation belongs to the cluster with the nearest centroid. In our experiments we used the free library by David Arthur, which implements an improved version called K-Means++ [5].

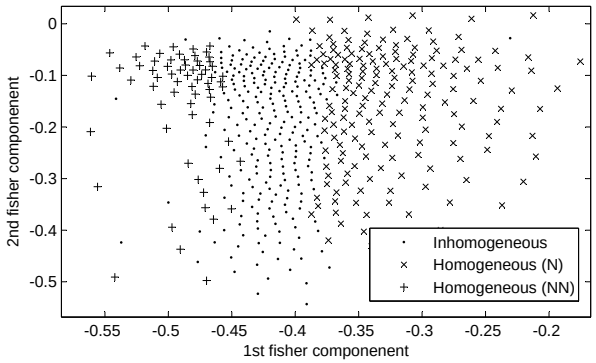

Figure 2: Distribution of *homogeneous* and *inhomogeneous* clusters for NEUROTICISM (*K=500* and *T=70%*). *Homogeneous* clusters are found at the right and left edges, *inhomogeneous* clusters accumulate in the center, where classes overlap.

For each cluster we then calculate the class distribution among its members. We do this by counting the frequency of frames belonging to the same class, divided by the total number of frames inside the cluster. Our assumption: the higher the relative amount of frames belonging to the same class, the more meaningful information is carried by the members of the cluster. We denote such clusters as *homogeneous*. Clusters that show a more or less equal class distribution, on the other hand, are not likely to host observations significant for any of the classes. Such clusters are regarded as *inhomogeneous*.

For the following pruning step we define: *If at least T % of the observations of a cluster are linked with class c, we regard the cluster as homogeneous with respect to c, otherwise we treat it as non-homogeneous.*

Figure 2 illustrates the result of that rule applied to the NEUROTICISM task ($K = 500$ and $T = 70\%$, i. e. a

cluster is regarded as *homogeneous* if at least 70% of the frames inside the cluster belong to the same class).

2.3. Pruning

After dividing the features space in *homogeneous* and *inhomogeneous* regions, we decide for every frame, whether to keep it for the classification process or not. This manifests the actual pruning step. Therefore, we simply determine for each frame its nearest cluster. Frames belonging to a *homogeneous* cluster are kept, all other frames are discarded. Note that this requires transformation of the frames into the previously defined Fisher space. But once the decision to keep a frame has been taken, we proceed with the original 129 features.

During our experiments, we found that longer sequences of *homogeneous* frames were sometimes interrupted by only a single or few *inhomogeneous* frames. To prevent such sequences from fragmentation we introduced a smoothing parameter N. If applied each frame decision, i. e. *homogeneous* or *inhomogeneous*, is replaced by a majority vote among its N neighbors. An additional parameter M defines the minimum length of an *inhomogeneous* block to be discarded. In combination with the threshold parameter T (see Section 2.2) it is now - at least to some extent - possible to control the amount of frames that will be discarded during pruning.

2.4. Classification

Finally, we prepare pruned sequences for final classification by extracting a set of high-level features (HLF). Again we stick to the procedure proposed by the organisers of the challenge and use the openSMILE tool [3] to extract a final set of 6125 features (see Table 5 in [2] for a listing). However, instead of raw wave files we input the pruned LLF sequences and remove components concerned with the calculation of low-level descriptors.

As classification scheme we employ Support Vector Machines (SVM) with Sequential Minimal Optimisation (SMO) and a polynomial kernel as provided by the WEKA data mining toolkit [6][2]. If we have successfully carved our cues that are distinctive for the classes, it should be easier for the classifier to distinguish them and this should lead to an improvement in recognition accuracy.

3. Results and Discussion

Of particular interest for our study is the question whether the proposed frame pruning algorithm actually leads to better recognition accuracies and how the amount of pruned frames connect to it.

[2]In particular, we chose WEKA's `functions.SMO` with `functions.supportVector.PolyKernel` and `C=0.01`, while keeping default settings for the remaining options.

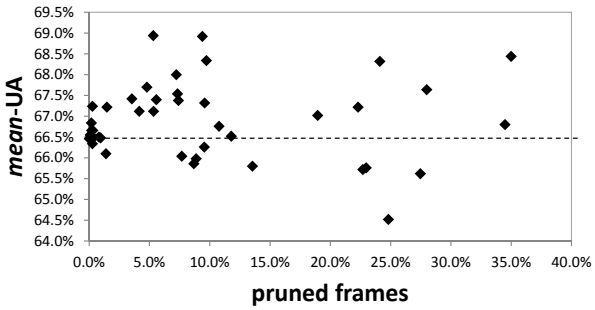

Figure 3: Amount of pruned frames in relation to `mean`-UA recall for investigated configurations. Improvements seem to be most stable within a range of 4%-8%. For higher pruning rates results increasingly diverge.

In search for an answer, we applied our approach to the Speaker Personality Corpus (SPC) [7] provided by the organiser of the INTERSPEECH 2012 Speaker Trait Challenge [2]. The corpus includes 640 clips from 322 individuals, most at a length of 10 seconds[3], labeled in five categories: OPENNESS, CONSCIENTIOUSNESS, EXTRAVERSION, AGREEABLENESS, NEUROTICISM. Each sample is connected to a set of 5 labels that denote which categories $X = \{O, C, E, A, N\}$ apply. If a category does not apply it is marked with a leading N. Samples are provided in three sets: a training set (n=256) to train the algorithm, a development (n=183) for evaluation, and an unlabelled test set (n=201) to participate in the challenge.

Due to the novelty of the approach it was difficult to predict good parameters from scratch. Hence, we decided to test a large number of configurations to learn about the individual characteristics of each parameter. During the experiments we used the proposed training set to determine the *homogeneous* clusters and pruned the development set accordingly.

After running some preliminary trials we decided to test combinations of the following set of parameters: $T = \{70\%, 85\%\}$, $K = \{500, 1000, 5000\}$ and $M = \{0s, 0.25s, 0.5s, 1.0s\}$. We also observed the influence of smoothing (see Section 2.3). If smoothing was applied ($smooth = true$) window size N was set to the same length as M, i. e. the minimum length of a *inhomogeneous* sequence to be discarded. Each combination of $\{T, K, M, smooth\}$ was evaluated with respect to the unweighted average (UA) recall measured for each individual task, as well as, the mean of all tasks (*mean*-UA). To set up a fair baseline we also included the special case where no frames are pruned[4].

[3]The relatively long sample duration suits frame pruning in so far that it makes it more likely that personality is not expressed with the same intensity throughout the whole sample.

[4]In the case where no frames are pruned extracted features should be identical with the feature set provided by the challenge organisers. However, we observed differences for some feature types. A reason

	UA recall in %						
	FP-NON	**FP-AVG**	**FP-BOC**	**FP-BIC**		**BASELINE**	
	devel	*devel*	*devel*	*devel*	*test*	*devel*	*test*
O	56.3	57.1	62.1	62.6	60.1	60.4	57.8
C	70.1	71.3	71.3	75.8	77.6	74.5	80.1
E	80.4	79.8	82.0	83.6	79.2	80.9	76.2
A	60.4	60.5	61.7	63.0	56.1	67.6	60.2
N	65.1	65.7	67.6	71.4	63.0	68.0	65.9
mean	**66.5**	**66.9**	**68.9**	**71.3**	**67.2**	**70.3**	**68.0**

Table 1: Recognition results: **FP-NON**=no pruning, **FP-AVG**=average among all configurations, **FP-BOC**=best overall configuration, **FP-BIC**=best individual configurations, **BASELINE**=challenge baseline.

In Figure 3 *mean*-UA for each configuration is plotted against the amount of pruned frames. We see that the majority of configurations exceed the baseline for the non-pruning configuration (dotted line), in the best case about 2.5%. The scatter plot also shows that stable improvement independent of the configuration is achieved within a moderate pruning range of 4%-8%. For higher pruning values results increasingly diverge and rates below the baseline become more likely. However, even for cases, in which more than 20% of the frames are dropped, some configurations achieve improvements of up to 2%.

Table 1 lists average result for all configurations, as well as, for the best single configuration $\{70\%, 1000, 1s, true\}$ and best individual configurations. In average the tested configurations cause a marginal improvement of 0.4%. However, if we pick a suited configuration, rates increase by 3-4%. In comparison to the challenge baseline we can report an improvement of 1% for the development set. However, the difference is not significant according to a McNemar's chi-squared test ($p > 0.05$). For the test set average accuracy is 0.8% below the baseline, due to a generally lower performance for the AGREEABLENESS class.

4. Conclusion

In this paper we have explored the effect of frame pruning on the INTERSPEECH 2012 Speaker Trait Challenge. A novel cluster-based algorithm was introduced, which aims at finding frames likely to carry distinctive information and keeps only them for the classification step. Results prove that under certain configurations frame pruning leads to an improved recognition accuracy. Most stable improvements were observed for a frame drop of 4-8%.

An advantage of the proposed algorithm is its universal applicability. Theoretically, it can be applied to any recognition problem that involves feature extraction on frame-level. Its application seems also promising in situations with varying noise and rest-class problems, as the algorithm is capable of isolating noisy or unfitting frames and excludes them from the classification process.

Several adaptations of the algorithms can be thought of. For instance, we must not necessarily operate on a fixed frame length, but may employ dynamic units, e. g. unvoiced vs. voiced chunks. An interpolation step may be added to get rid of sharp edges where frames have been dropped. Finally, instead of keeping all $homogeneous$ frames we could keep only those that are $homogeneous$ with respect to a certain class and learn an ensemble of classifiers, one for each class. Final decision can be obtained by fusing results from the individual models.

An implementation of the proposed algorithm in Matlab is available from `http://hcm-lab.de/is12.html`.

could be that some meta information, such as the sample rate of the input signal, differs when high-level features are extracted from pre-calculated low-level features instead of the raw wave files.

5. Acknowledgements

The presented work is funded by the EU under research grant CEEDS (FP7-ICT-2009-5), ILHAIRE (FP7-ICT-2009.8.0) and TARDIS (FP7-ICT-2011-7).

6. References

[1] L. Bessacier and J. Bonastre, "Frame pruning for speaker recognition," in *ICASSP98*, 1998, pp. 765–768.

[2] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, f. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The interspeech 2012 speaker trait challenge," in *in Proceedings of INTERSPEECH*, 2012.

[3] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 1459–1462.

[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. John Wiley & Sons Inc, 1973.

[5] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[7] G. Mohammadi, A. Vinciarelli, and M. Mortillaro, "The voice of personality: Mapping nonverbal vocal behavior into trait attributions," in *Proceedings of ACM Multimedia Workshop on Social Signal Processing*, 0 2010.