Towards Measuring Similarity Between Emotional Corpora

Mátyás Brendel, Riccardo Zaccarelli, Björn Schuller, Laurence Devillers

LIMSI-CNRS
Spoken Language Processing Group
91403 Orsay, France

mbrendel, riccardo, schuller, devil@limsi.fr

Abstract

In this paper we suggest feature selection and Principal Component Analysis as a way to analyze and compare corpora of emotional speech. To this end, a fast improvement of the Sequential Forward Floating Search algorithm is introduced, and subsequently extensive tests are run on a selection of French emotional language resources well suited for a first impression on general applicability. Tools for comparing feature-sets are developed to be able to evaluate the results of feature selection in order to obtain conclusions on the corpora or sub-corpora divided by gender.

1. Introduction

At present, there are various corpora in use by the automatic emotion recognition community, with considerable difference in size, topic and application context. None of them are ideal: they all have their advantages and drawbacks. Consequently, the comparison and possible unification of corpora is an important aid for current research: (Tahon and Laurence Devillers, 2010) for example studies differences in anger across corpora in a very detailed fashion by examining acoustic properties. In this paper we take the same two corpora in order to allow for meaningful comparisons and findings, yet with different methods. We will be using PCA and feature selection to visualize and compare corpora by their compound of most relevant features. This seems reasonable, as it is e. g. known that 'more' acted corpora tend to prefer pitch-based descriptors in comparison to more natural emotional speech, where spectral information typically is expected as 'reliable candidate' of best features.

Feature selection is usually considered as a tool to make machine learning models more efficient. Selecting the best features may improve the quality of the model by avoiding over-training and/or it may increase the speed of computation and reduce memory demands. Feature selection, however, is rarely considered as a tool to characterize or analyze a corpus or measure similarities or differences among corpora.

Sequential Floating Forward Selection (SFFS) was introduced in (Pudil et al., 1994) and is certainly among the most widely used techniques in the field. In this paper, we consider SFFS as the baseline experimental technique and try to improve it by the help of 'set-similarity'. This approach is different to other improvements found in the literature, and it introduces a different dimension to amend feature selection algorithms.

As named above, our intent is to analyze the similarities and differences of corpora or sub-corpora by their compound-structure of highly relevant features. The main motivation of introducing a modified variant of SFFS is the comparably high computation cost of SFFS. This is a clear drawback in the respect of our aim: if a feature-selection-based comparison of several corpora is to be carried out, there will be a clear demand for sufficient speed of processing. By using our proposed algorithm a more extensive analysis is possible as it would be achievable in the same amount of time using 'classical' SFFS.

CINEMO	# POS	# SAD	# ANG	# NEU
# segments	313	364	344	510

Table 1: CINEMO sub-corpus, number of segments for 50 speakers

2. Corpora

The corpora CINEMO and JEMO were already introduced in (Brendel et al., 2010). Here we only give a short description.

2.1. CINEMO

The corpus CINEMO (Rollet et al., 2009) used in this paper consists of 1 532 instances after segmentation of emotional French speech amounting to a total net playtime of 2:13:59 hours. 50 speakers (of 15 to 60 years old) dubbed 27 scenes of 12 movies. A subset of the more consensual segments was chosen for training models for detection of 4 classes (POsitive, SADness, ANGer and NEUtral). The rich annotation of CINEMO was used to build these 4 macro-classes. Table 1 shows the distribution of instances among classes within the considered CINEMO sub-corpus.

2.2. JEMO

The corpus JEMO features 1 135 instances after segmentation of speech recorded from 39 speakers (18 to 60 years old). JEMO is a corpus collected within an emotion detection game. This game used a segmentation tool based on silenced pauses and used a first system of 5 emotions detection (ANGer, FEAr, SADness, POsitive and NEUtral) and a system of activation detection (low/high) built on CINEMO data. The corpus recorded was the reaction of the users to the system response.

JEMO	# POS	# SAD	# ANG	# NEU
# segments	316	223	179	416

Table 2: JEMO sub-corpus, number of segments for 39 speakers.

C. & J.	# POS	# SAD	# ANG	# NEU
Male	252	262	267	432
Female	377	325	256	494

Table 3: Female and Male sub-corpora of the unified corpus, # of segments for 38 female and 50 male speakers.

In JEMO speakers generated spontaneous sentences with higher level of expressivity than in CINEMO. The corpus has been annotated by two coders with major and minor emotions. These data were more prototypical than in the corpus CINEMO as very few mixtures of emotions were annotated.
Table 2 shows instance distribution in the JEMO sub-corpus. Table 3 shows instance distribution in the sub-corpora obtained by dividing the unified corpus by gender.

3. Features

In the following we will describe two different feature sets based on two different extraction engines.

3.1. LIMSI features

Each speech segment is passed through spectral (16 MFCCs) and prosodic analysis (pitch, zero-crossing and energy) by the LIMSI extractor. The feature extractor next calculates basic statistical features on voiced parts: min, max, mean, standard deviation, range, median quartile, third quartile, min and max intra range and the mean and standard deviation of the coefficients of least square fitting regression (of each voiced segment); min and max inter range (between voiced segments). Overall, 458 features are thus obtained including further post-processing: 23 for pitch, 51 for energy (from these 22 root mean square energy), 18 zero-crossings and 366 for MFCC1–16.
Table 4 shows the low level descriptors and functionals used in generating the LIMSI features for these experiments.

LLD	Functionals
Energy	*moments(2):*
RMS Energy	absolute mean, max
F0	*extremes(3):*
Zero-Crossing-Rate	2 x values, range
MFCC 1–16	*linear regression(2):*
	MSE, slope
	quartiles(2)
	quartile, tquartile

Table 4: *LIMSI features: low-level descriptors and functionals. Abbreviations: root mean square (RMS), Mel Frequency Cepstral Coefficients (MFCC), Mean Absolute/Square Error (MAE/MSE). Note that not all combinations are used.*

LLD	Functionals
(δ) RMS Energy	*moments (4):*
(δ) Log-Frame-Energy	absolute mean, std. deviation
(δ) Voicing Probability	kurtosis, skewness
(δ) F0	*extremes (5):*
(δ) F0 envelope	$2 \times$ values, $2 \times$ position, range
(δ) Zero-Crossing-Rate	*linear regression (4):*
(δ) MFCC 1–12	offset, slope, MAE, MSE
(δ) LSP Frequency 0–7	*quartiles (6):*
	$3 \times$ quartiles, $3 \times$ ranges

Table 5: *Acoustic features in openEAR: low-level descriptors and functionals. Abbreviations: Line Spectral Pairs (LSP), Mel Frequency Cepstral Coefficients (MFCC), Mean Absolute/Square Error (MAE/MSE).*

3.2. openEAR features

To introduce sufficient variance in our experimentation and not base our findings solemnly on one feature extractor, we use the same openEAR toolkit's (Eyben et al., 2009) *"base"* set as used in (Schuller et al., 2010): 988 features – a slight extension over the set provided for the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009) – based on 19 functionals of 26 acoustic low-level descriptors (LLD, smoothed by simple moving average) and corresponding first order delta regression coefficients as depicted in Table 5.

4. Using PCA for visualizing corpora

To illustrate the distribution of the corpus, the mean of each feature was computed per speaker and per class. In order to be able to display the speaker-means the most important principal components were computed with Weka (Witten and Frank, 2005) and the first two components are shown.
Figure 1 shows the speaker-means of CINEMO with 458 LIMSI features in the first two dimensions of its PCA space.
Figure 2 shows the speaker-means of JEMO with 458 LIMSI features in the first two dimensions of its PCA space.
Figure 3 shows the speaker-means of CINEMO with openEAR features in the first two dimensions of its PCA space.
Figure 4 shows the speaker-means of JEMO with openEAR features in the first two dimensions of its PCA space.
Comparing these figures one can see that the LIMSI features form an elongated shape, especially in JEMO, while with openEAR the shapes are rounder, which seems to be better. Although classes are intertwingled in all the cases, the openEAR figures show more separation of the classes, especially on JEMO. Consequently, we can expect better results with JEMO and with openEAR features. This will be confirmed in the following sections.
PCA is a dimension-reduction technique suited to display instances of a high-dimensional space in a lower dimensional one. However, the principal components are not easily interpretable for humans. It seems worth, though, to select 2 important features and compare the two sub-corpora in this two dimensional and interpretable space.
Figures 5 and 6 show the classes ANG and POS of CINEMO and JEMO in the feature-space of MeanEnergy and MeanPitch. In both cases we can see some differences between

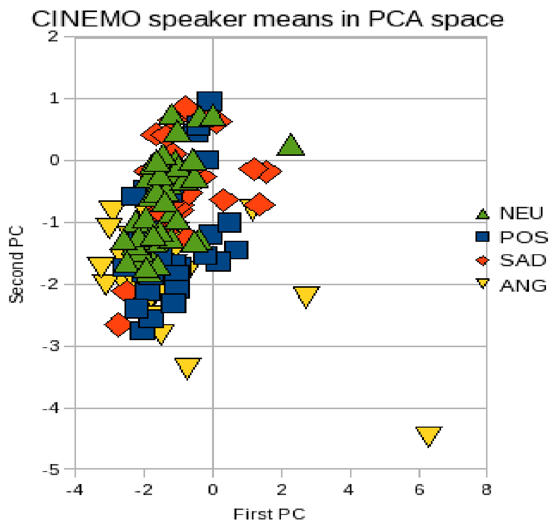

Figure 1: Speaker-means of CINEMO with 458 LIMSI features in the two most important dimensions of its 2D PCA space. Note that some data of all the classes is masked by the classes NEU and POS in the center.

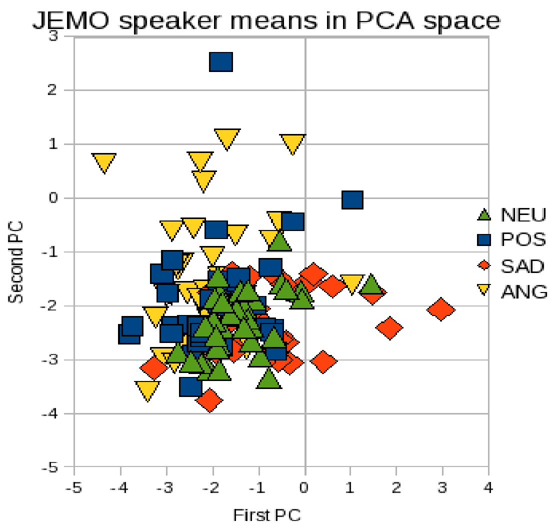

Figure 2: Speaker-means of JEMO with 458 LIMSI features in the first two dimensions of its 2D PCA space. Note that some data of all the classes is masked by the classes NEU and POS in the center.

the sub-corpora. Anger in JEMO often contains more energy and higher pitch than in CINEMO. Also in the case of POS of JEMO sometimes higher energetic levels are observed. Note that the coordinates of the two figures are not the same, i. e. the energy related to POSitive is usually lower than that related to ANGer.

5. Using feature selection for measuring differences between corpora

A good introduction to feature selection can be found in (Guyon and Elisseeff, 2003). Methods are generally divided into two larger groups: filter-based and wrapper methods.

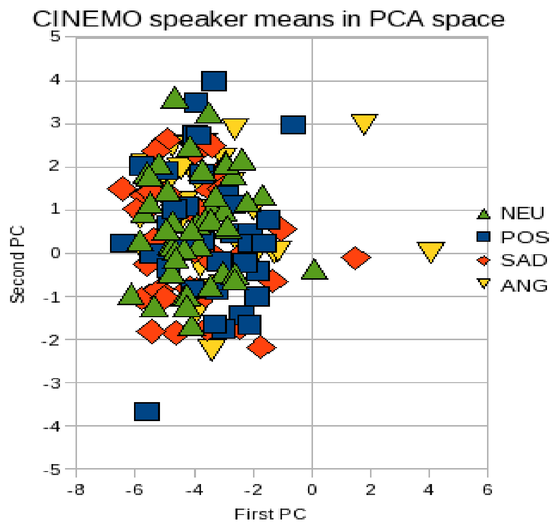

Figure 3: Speaker-means of CINEMO with openEAR features in the two most important dimensions of its 2D PCA space. Note that some data of all the classes is masked by the classes NEU and POS in the center.

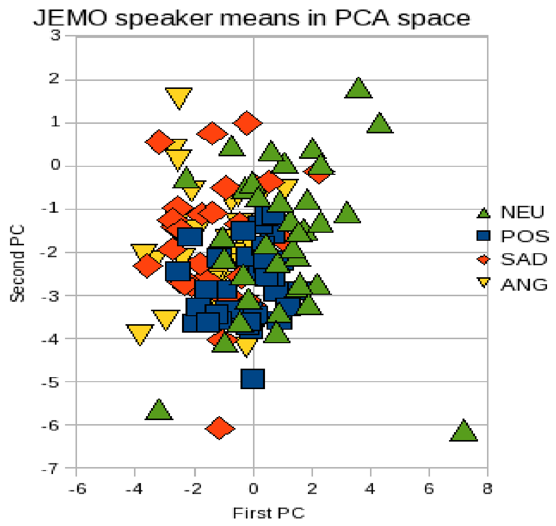

Figure 4: Speaker-means of JEMO with openEAR features in the first two dimensions of its 2D PCA space. Note that some data of all the classes is masked by the classes NEU and POS in the center.

Filter-based variable ranking is usually computationally affordable, since often only a simple scoring function is computed. However, it usually can not take into account the interaction or correlation between features. At the same time even 'weak' features may add considerably in a compound and should thus not be discarded by choosing only individually high ranked candidates. Wrappers utilize a data-driven learnt classifier's minimal error as target function – consequently they are time-consuming once more complex algorithms are chosen. Usually, one would like to have the later target classifier also employed in the selection process to avoid biases. In this paper we provide results with a wrapper method, namely Sequential Forward Float-

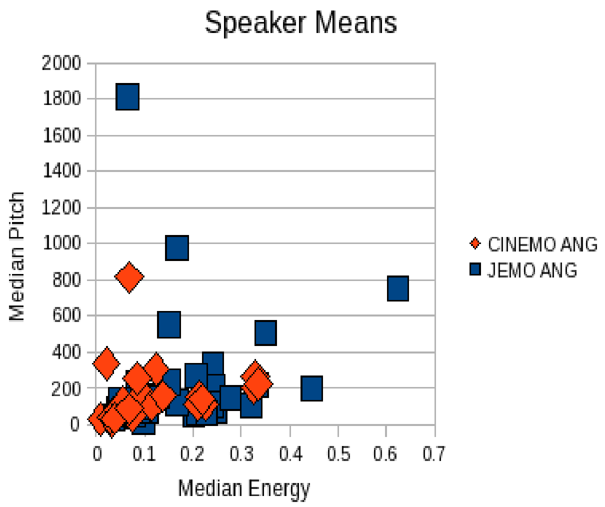

Figure 5: Speaker-means of class ANG in JEMO and CINEMO in the feature-space of MeanEnergy and MeanPitch.

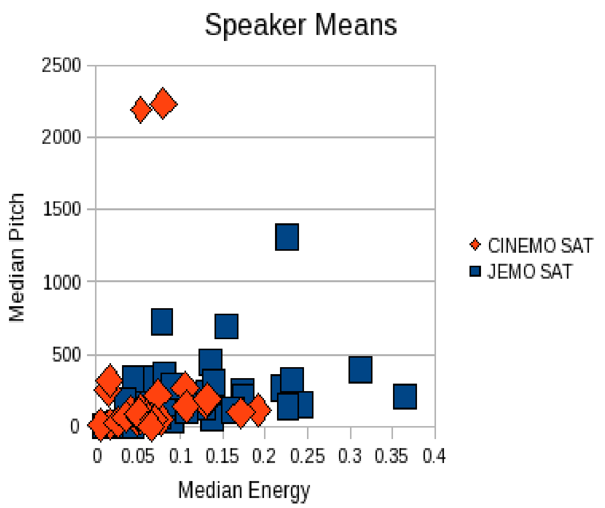

Figure 6: Speaker-means of class POS in JEMO and CINEMO in the feature-space of MeanEnergy and MeanPitch.

ing Search (SFFS) and alternatives that are computationally less 'expensive'.

5.1. Improving feature selection with set-similarity based heuristics

Let $Y = \{y_i : 1 \leq i \leq D\}$ denote a set of available features and $X = \{x_i : 1 \leq i \leq k, x_i \in Y\}$ a subset of features. The named wrapper-methods do feature selection by running a real test on the subset X. In our case this test is always a 10-fold speaker independent (SpI) cross-validation (CV). As described, the worth is evaluated by a measure – in our case recognition rate – and the result of this evaluation is denoted by $J(X)$, i. e. J is our criterion function.
Feature selection can usually be considered as a tree-search, especially when the branch and bound method (B&B) is considered (Somol et al., 2004). Since we consider forward selection, our tree is reverted compared to B&B: the root node represents the empty set and the child nodes of a parent represent all possible extensions of the parent's feature-set with one new feature. The root node has D children, and the number of children is exactly $D - k$ on level k.
As in (Somol et al., 1999), the forward step – the Sequential Forward Search (SFS) – is similar to a breadth-first search on this tree with the rule that on each level only the optimal child is selected while all others are pruned. In addition, there is a backward step, in which a feature is removed, if a better feature-set is obtained measured by the defined optimization criterion than the so far optimal one on that level.
Feature selection can also be considered as a global optimization task, and the forward step can be considered as exploitation and the backward step as exploration. In global optimization the balance of exploration and exploitation is important: without the backward step, SFFS would be too greedy: it would too easily be stuck in local optimums. However, what we will show in the ongoing is that the backward step is strong enough to make the forward step more greedy. At a certain level k, SFFS tests all the $D - k$ nodes and takes the best. A more greedy algorithm would be more similar to a depth-first search, meaning that we take the first child with a positive gain. This way exploitation is made faster, which however, increases the danger to stick to a local optimum. Nevertheless, our claim is that in our field of application, the backward step is strong enough to handle this.
A further improvement was made to order the new feature candidates according to the expectation of their significance. To estimate the gain achievable by adding a certain feature x to the set of features X, we use the known history of our search tree: we take the most similar case, when x was added to a feature set X'. We will call the gain for x the significance of x: $S(x, X)$, which is similar to the notation of (Somol et al., 1999). The estimated significance is denoted as $S'(x, X)$.
Similarity of two sets may be measured in many ways, one of the most frequently used measures is Jaccard similarity, which we decided for. This means that estimated significance is computed as follows:

$$\begin{aligned} S'(x, X) &= J(X^* \cup \{x\}) - J(X^*) : \\ X^* &= argmax(Jaccard(X, X')) \end{aligned} \tag{1}$$

where $Jaccard()$ is the Jaccard set similarity measure. Instead of $argmax$, other functions may be used, like for example a weighted sum with exponentially decaying weights. In the backward step of SFFS we keep the breadth-first manner of the algorithm, since a strong exploration is needed to avoid local minima. However, not a full breadth search is done, only a certain percentage p of the ordered candidates are tested. Candidates are ordered in increasing order of estimated significance, so that we try to remove first the most insignificant features. We applied p=20 % of breadth search, which proved to be sufficient in our case.
We name our introduced method "SFFS with Set-Similarity Heuristics" (SFFS-SSH). This proposed algorithm has only one parameter to be tuned: p. Note that the computational overhead of our heuristics is negligible compared to a 10-fold CV test on a corpus. Thus, the running time of the

# it. / RR	SFFS	SFFS-SSH	all feat.
LIMSI	28 382 / 53.4 %	6 394 / 52.4 %	– / 54.6 %
openEAR	28 126 / 58.5 %	4 742 / 58.8 %	– / 59.6 %

Table 6: Number of iterations and Recognition Rate (RR) using the best 24 selected features on the united CINEMO and JEMO corpus with conventional SFFS in comparison to our suggested efficient modification (SFFS-SSH).

RR	24 L.	24 O.	all L.	all O.
CINEMO	49.3 %	56.6 %	48.5 %	53.8 %
JEMO	63.7 %	67.9 %	61.6 %	64.2 %
Female	64.6 %	64.5 %	59.5 %	62.6 %
Male	53.0 %	56.6 %	49.5 %	57.0 %

Table 7: Recognition Rate (RR) with the best 24 selected features, and all features. Abbreviations: openEAR (O.), LIMSI-features (L.)

LIMSI	MFCC	Pitch	Energy	ZCR
CINEMO	21	0	2	1
JEMO	21	0	2	1
Female	17	3	3	1
Male	17	5	2	0

Table 9: Frequency of different feature groups in the 24 selected LIMSI features.

openEAR	MFCC	Pitch	Energy	Zrc
CINEMO	12	0	6	1
JEMO	11	5	2	0
Female	10	3	3	4
Male	6	1	8	0

Table 10: Frequency of different feature groups in the 24 selected openEAR features. Note that if features are missing to sum up to 24, they are of other kind than the considered ones.

algorithm depends mainly on the number of iterations, which we expect to decrease significantly.

In the following experiments we trained the data set using LIBSVM (Chih-Chung Chang and Lin, 2001) with a radial basis function kernel. As stated earlier, we use 10-fold speaker independent cross-validation, designed in our lab. In short this means that speakers are divided into folds, instead of partitioning merely taking instances without speaker assignment into account, thus maintaining speaker independence, while being able to run a 10-fold CV with all its benefits as being able to use a complete (sparse) data set for testing and introduce variance in the evaluative runs.

Table 6 shows the results with a fixed number of 24 features. As can be seen, selecting 24 features does not improve the result. The explanation for this is that the united corpus is sufficiently large to avoid over-training. It can further be seen that SFFS-SSH provides similar good feature-selection at considerably lower number of iterations. Thus the first test of our method was successful, which is why we will exclusively apply SFFS-SSH in the ongoing.

Table 7 shows the recognition rate (RR) for the various sub-corpora with SFFS-SSH with the number of features fixed to 24.

There is a more significant difference in RR between CINEMO and JEMO than between the female and male sub-corpora for both libraries (openEAR and LIMSI) and also for the selected features and the total set of features.

Table 8 next shows the recognition rate (RR) for the various sub-corpora with SFFS-SSH with the optimal number of features.

RR / # features	LIMSI	Openear
CINEMO	55.7 % / 29	58.2 % / 36
JEMO	65.8 % / 43	72.2 % / 43

Table 8: Recognition Rate (RR) and number of features of the best selected feature set

As seen in the table, some percent of improvement can be achieved with a significantly higher number of features. Interestingly, the optimal number of features is higher for JEMO with both feature-sets.

5.2. Ratio of feature groups in selected feature sets

Having established an efficient method for feature selection we will next consider how it can be used for our primary aim in this paper: in order to measure the differences between sub-corpora, as a first attempt we have computed the ratio of different feature groups for LIMSI and openEAR after feature selection.

Tables 9 and 10 show the number of features in the different features groups. Since the number of features is constantly 24, these correspond to ratios. There are considerably less MFCC features used in openEAR in overall ratio, but more corresponding to energy and other low-level descriptors, not grouped here (cf. table 5).

The differences between female-male seems to be larger than between CINEMO and JEMO. This contradicts the difference in RR. Consequently, these numbers are important but not detailed enough. We thus will next investigate further measures.

Since CINEMO and JEMO in table 9 have exactly the same ratios for each feature group, we repeated this experiment with 48 features. This naturally demands for longer computation times, as SFFS is a forward selection. Thus, it would have been desirable to reveal differences already at a low dimension of the selected feature space. For quantitative illustration we consider this experiment with the LIMSI set on the CINEMO and JEMO corpora. Results are shown in table 11.

Visibly, there is a slight difference in this case compared to the smaller target set size, but results across corpora are very similar. Comparing table 9 to table 11, we can only see that more energy features have been selected, which likely indicates their lower relevance, though used, if more features are to be selected. Recognition rate is 64.6 % for

LIMSI	MFCC	Pitch	Energy	ZCR
CINEMO	32	1	13	2
JEMO	30	1	16	1

Table 11: Number of different feature groups in the 84 selected LIMSI features.

Similarity of features	LIMSI	openEAR
CINEMO–JEMO	0.5983	0.5903
Female–Male	0.4907	0.4255

Table 12: Correlation-based similarity of the selected feature-sets.

JEMO and 51.1 % for CINEMO, which resembles a slight improvement.
This extended experiment did not bring us further – it just confirmed previous results: we consider more features and by that obtain slightly improved results in terms of recognition rate, however, we still need tools for more detailed analysis of the features.

5.3. Correlation based similarity of feature-sets

Having two feature sets from the same total set of features, one would like to compare the two sets. To list the features selected appears to have less practical applicability, since there might be different features, which are similar. For the same reason, a simple Jackard-similarity is also not sufficient. Instead, measures of describing and comparing feature-sets have to be developed.
One way to measure similarity of corresponding features is the cross-correlation matrix. There is no way to define a cross-correlation between the same features of the different sub-corpora, like CINEMO and JEMO or female and male, since the instances are independent. What can be done, though, is to compute a cross-correlation for different features over the united corpus, i. e. for CINEMO and JEMO together. Moreover, we can define the similarity of the selected feature-sets based on this. The similarity of feature sets F and F' is computed as follows:

$$sim(F, F') = aver\{f \in F : min\{corr(f, f') : f' \in F'\} \quad (2)$$

Where "aver" is the average and "corr" the correlation computed on the entire corpus. Since the measure "sim" is asymmetrical, the average of $sim(F, F')$ and $sim(F', F)$ was taken. Note that this measure would not have any sense over the totals of features (it would be 1) – it is only reasonable in combination with feature selection.
In table 12 the similarity of sub-corpora can be seen measured by the similarity of the 24 selected features. As can be seen, the similarity of CINEMO and JEMO in terms of feature sets is higher than of female to male. This confirms our finding in recognition rate.

5.4. Rank based similarity of feature-sets

It is not straight forward to derive an individual feature's relevance in the resulting feature-set. To obtain a sharpened

Difference in feature ranks	LIMSI	openEAR
CINEMO-JEMO	0.17	0.19
Female-Male	0.07	0.09

Table 13: Highets differences in feature ranks.

picture on this, one can use the first iteration of SFFS as a feature ranking, as it considers each individual. Note that the first iteration of SFFS and our SFFS-SSH is identical. The rank of the feature is the result obtained using only that one feature. One can do this not only for the selected features but for the total set of features, i. e. 458 for LIMSI features and 988 for openEAR in our case.
Table 13 shows the highest difference in feature ranks for the same features. This measure shows that CINEMO and JEMO is more different in this aspect than the female and male sub-corpora. There is a consistently higher difference for openEAR, the reason might be that results for openEAR are better and the number of features is higher, which allows for the maximum to be higher. The measure based on feature ranks by that confirms the measure based on feature-groups.

6. Conclusions

We have seen four measures connected to feature selection for measuring similarity of sub-corpora: similarity measures based on recognition rate, groups of features, correlation and feature-ranks. They support however two different kinds of results: for recognition rate and correlation, the difference between female and male – which was considered as comparative anchor – is higher than between CINEMO and JEMO. On the other hand, measured in feature-groups and feature-ranks, the difference between CINEMO and JEMO is higher than the difference between female and male. Our result indicates that several measures have to be used. The four measures seem also to show that there might be at least two aspects of difference: in one aspect female and male are more different, in another the CINEMO and JEMO sub-corpora are.

7. Future Work

For dimension reduction, future work may test other techniques, especially so called cluster-preserving or similarity-preserving transformations.
In the future several more measures shall be developed to measure similarity of feature sets and corpora. Similarity of feature sets and the importance of features might be represented in a tree-like structure, which corresponds to the tree-structure of feature selection and propagates some measures of importance and similarity. This structure and measure is however complex.
Our SFFS-SSH algorithm can also be further developed. For example, instead of Jaccard-similarity a correlation based similarity measure may be used not only after feature selection but already within feature selection.
The correlation-based similarity measure can also be improved: for each feature currently we take only into account the most similar feature. A more complex measure would add together all the similar features with a decreasing weight.

8. References

M. Brendel, R. Zaccarelli, and L. Devillers. 2010. Building a system for emotions detection from speech to control an affective avatar. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Ch.-Ch. Chih-Chung Chang and Ch.-J. Lin, 2001. *LIBSVM: a library for support vector machines*.

F. Eyben, M. Wöllmer, and B. Schuller. 2009. openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *Proc. ACII*. IEEE.

Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

P. Pudil, J. Novovičová, and J. Kittler. 1994. Floating search methods in feature selection. *Pattern Recogn. Lett.*, 15(11):1119–1125.

N. Rollet, A. Delaborde, and L. Devillers. 2009. Protocol cinemo: The use of fiction for collecting emotional data in naturalistic controlled oriented context,. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction, 2009*.

B. Schuller, S. Steidl, and A. Batliner. 2009. The INTERSPEECH 2009 Emotion Challenge. In *Proc. Interspeech*, pages 312–315, Brighton, UK. ISCA.

B. Schuller, R. Zaccarelli, N. Rollet, and L. Devillers. 2010. Cinemo a french spoken language resource for complex emotions: Facts and baselines. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

P. Somol, P. Pudil, J. Novovičová, and P. Paclík. 1999. Adaptive floating search methods in feature selection. *Pattern Recogn. Lett.*, 20(11-13):1157–1163.

P. Somol, P. Pudil, F. J. Ferri, and J. Kittler. 2004. Fast branch & bound algorithms for optimal feature selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(7):900–912.

M. Tahon and L. Laurence Devillers. 2010. Acoustic measures characterizing anger across corpora collected in artificial or natural context. In *Proceedings of the Fifth International Conference on Speech Prosody, 2010*.

I. H. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.